Alma Mater Studiorum · Università di Bologna

SCUOLA DI SCIENZE MATEMATICHE, FISICHE E NATURALI
Corso di Laurea Magistrale in Matematica

METODI MATEMATICI PER LO SVILUPPO DI UNA NUOVA DISTANZA GENETICA PER INFERIRE EVENTI DEMOGRAFICI DA DATI DI SEQUENZIAMENTO DI POPOLAZIONI UMANE

Tesi di Laurea in Analisi di Dati

Relatore:
Chiar.mo Prof.
FAUSTO DESALVO

Presentata da:
GIULIA ZEPPILLI

Correlatori:
MASSIMO CAMPANINO
ALESSIO BOATTINI
LUCA PAGANI

II Sessione
Anno Accademico 2013-2014

La varietà è l'essenza del sapore.
Proverbio zanzibariano

A Domenica, Osvaldo,
Maria e Fernando

Introduzione

La presente trattazione nasce dall'applicazione di alcuni metodi matematici a ricerche e analisi di dati di carattere antropologico, cioè relativi a popolazioni umane.
Si apre con un capitolo in cui vengono descritti, da un punto di vista prettamente teorico, i metodi matematici utilizzati per l'analisi dei dati e i principi su cui essi si basano, in modo da permettere di comprendere le motivazioni che hanno spinto all'utilizzo di tali metodi. Successivamente viene data una breve spiegazione delle nozioni di antropologia e genetica delle popolazioni, necessarie alla comprensione e all'interpretazione dei risultati del lavoro svolto. Nel terzo capitolo vengono descritti i materiali a disposizione per le ricerche, i dati e ripresi i metodi del primo capitolo ma da un punto di vista pratico e strettamente collegato all'utilizzo nell'ambito della ricerca. Nel quarto capitolo sono presentati i risultati, i grafici, le tabelle e tutto quello che è stato ottenuto dalle analisi. Infine, nel quinto capitolo, verranno spiegati e discussi i risultati per interpretarli e trarne conseguenze utili all'antropologia.

Indice

Capitolo 1

I metodi matematici utilizzati

I metodi matematici alla base della presente trattazione rientrano principalmente negli strumenti della statistica e probabilità e dell'analisi numerica e verranno descritti in questo capitolo.

In particolare, per quanto riguarda la statistica, verranno descritte alcune tecniche di statistica multivariata dei dati, che permettono di analizzare simultaneamente misurazioni riguardanti diverse caratteristiche di un insieme di individui in esame, come di fatto avviene spesso in ambito scientifico e come si è fatto nella trattazione. Gli obiettivi principali di tali metodi sono sostanzialmente la semplificazione della struttura delle osservazioni, il loro ordinamento e classificazione.

Verrà presentata, inoltre, una procedura di statistica inferenziale, che è una branca della statistica che induce le caratteristiche di un'intera popolazione dall'osservazione di un campione di essa, selezionato solitamente mediante un esperimento aleatorio, come si è proceduto anche nella tesi, mediante un programma di simulazione di popolazioni umane.

Si parlerà infine di simulazione, essendo stati di fondamentale importanza, nel corso del lavoro svolto, i dati relativi a popolazioni generate mediante un software di simulazione.

Per ciò che concerne invece l'analisi numerica, si approfondirà il concetto di interpolazione.

1.1 La Cluster Analysis

1.1.1 Introduzione alla classificazione

Una delle più basilari abilità delle creature viventi riguarda il raggruppare simili oggetti per produrne una loro classificazione; si tratta di un'idea primitiva: per esempio, già i primi umani hanno dovuto imparare a raggruppare oggetti che condividevano le stesse proprietà, come essere cibi edibili o velenosi, animali feroci o innocui.
In senso più ampio, la classificazione è stata necessaria per lo sviluppo del linguaggio, fatto di parole che aiutano a distinguere e discutere su eventi, oggetti, persone; semplificando, ogni parola non è altro che un'etichetta utilizzata per descrivere una classe di cose che hanno una caratteristica in comune: nominare e classificare sono essenzialmente sinomini.
La classificazione è anche fondamentale per molte brance della scienza. In biologia, ad esempio, la classificazione degli organismi è stata da sempre oggetto di preoccupazione. Fu Aristotele il primo ad elaborare un sistema di classificazione delle specie animali; fu seguito da Teofrasto, autore di una classificazione delle piante superata solo dagli studiosi e esploratori europei del XVII e XVIII secolo, quali Linneo e Darwin. In chimica, durante gli anni '60 dell'800, la classificazione degli elementi di Mendeleyev ha avuto un notevole impatto sulla compresioe della struttura dell'atomo. In astronomia, le teorie dell'evoluzione delle stelle hanno risentito della classificazione delle stelle in nane e giganti, condotta mediante il diagramma di Hertzsprung-Russel, il quale mette in relazione la temperatura effettiva (riportata in ascissa) e la luminosità (riportata in ordinata) delle stelle.

1.1.2 Metodi numerici di classificazione: la Cluster analysis

Le tecniche matematiche per la classificazione si sono originate in gran parte nelle scienze naturali, come la biologia e la zoologia, nel tentativo di liberare la tassonomia dalla sua natura soggettiva. Lo scopo era quello di fornire classificazioni oggettive e stabili.
Il termine che oggi raggruppa tutti questi metodi volti al raggruppamento di dati, utilizzati nelle più varie discipline, quali biologia, psicologia, marketing, e chiamati con nomi diversi da ognuna è *Cluster analysis*.
In molte applicazioni della Cluster analysis è richiesta una partizione dei dati, cioè ogni oggetto appartiene a un singolo raggruppamento (cluster) e l'insieme completo dei cluster contiene tutti gli oggetti. In alcuni casi invece,

è ammessa una sovrapposizione dei cluster, anzi è utile per una più accettabile soluzione. È necessario però ricordare anche che una risposta accettabile di una Cluster analysis può essere che nessun raggruppamento dei dati sia giustificato.

Sia data una popolazione π_I; di questa si considerino n individui $I = I_1, I_2, ..., I_n$. Si assume che esista un insieme di p caratteristiche, $C = (C_1, C_2, ..., C_p)^T$, osservabili (cioè che possano essere misurate) e possedute dagli individui di I. La misura dell'$i-esima$ caratteristica dell'individuo I_j si denota con x_{ij} mentre $X = [x_{ij}]$ denota il vettore $p \times 1$ di tali misure. Quindi, per un insieme di individui I, si dispone di un insieme di vettori $p \times 1$ di misure, cioè $X = X_1, X_2, ..., X_n$, che descrivono l'insieme I.
Sia m un intero strettamente minore di n. Basandosi sui dati contenuti in X, il problema della Cluster analysis è quello di determinare m cluster di individui, $\pi_1, \pi_2, ..., \pi_m$, in modo che l'individuo I_i appartenga a uno e un solo cluster e che sia simile agli individui assegnati allo stesso cluster e diverso (non simile) a quelli appartenenti ad altri cluster.
Per poter risolvere questo problema è necessario definire cosa significa *essere simili* in termini quantitativi, cioè cosa si intende dicendo che due individui I_i e I_k sono diversi o meno.
Una soluzione a questo problema potrebbe derivare dall'aver assegnato l'$i-esimo$ e il $k-esimo$ individuo allo stesso cluster se la distanza fra X_i e X_k è *sufficientemente piccola* e a cluster diversi se invece la distanza è *sufficientemente grande*. Dunque è nostro obiettivo capire anche cosa si intende per distanza tra X_i e X_j.

Una funzione reale non negativa $d : X \times X \longrightarrow \mathbb{R}$ si chiama distanza o metrica se:

- $d(x_i, x_j) \geq 0$;
- $d(x_i, x_j) = 0 \iff x_i = x_j$;
- $d(x_i, x_j) = d(x_j, x_i)$ (simmetria);
- $d(x_i, x_j) \leq (x_i, x_k) + d(x_k, x_j)$ (disuguaglianza triangolare)

$\forall x_i, x_j, x_k \in X$.

Il valore di $d(X_i, X_j)$ tra X_i e X_j si chiama distanza ed è equivalente alla distanza tra I_i e I_j rispetto alle caratteristiche $C = (C_1, C_2, ..., C_p)^T$ selezionate.

Esempi di distanze comuni sono:

- distanza Eucludea: $d_2 = \sqrt{\sum_{k=1}^{p}(x_{ki} - x_{kj})^2}$
- 1-distanza: $d_1 = \sum_{k=1}^{p}(x_{ki} - x_{kj})^2$
- p-distanza: $d_2 = \sqrt[p]{\sum_{k=1}^{p}(x_{ki} - x_{kj})^p}$.

Nelle applicazioni di Cluster analysis nel corso della trattazione è stata utilizzata sempre la distanza euclidea.

Dato l'insieme di osservazioni $X = X_1, X_2, ..., X_n$ degli individui $I = I_1, I_2, ..., I_n$, si indica con s_d lo scalare

$$s_d = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} d(x_i, x_j)$$

detto dispersione totale rispetto alla distanza $d(x_i, x_j)$.
Il valore $\bar{s}_d = \frac{s_d}{N_d}$, dove $N_d = \frac{n^2-n}{2}$, è detto invece dispersione media dell'insieme I.
Le definizioni dei valori s_d e $\bar{s}_d$ vengono dalla considerazione della matrice

$$D = d_{ij} = d(X_i, X_j) = \begin{bmatrix} 0 & d_{12} & \cdots & d_{1n} \\ d_{21} & 0 & \cdots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & 0 \end{bmatrix}$$

della quale si nota che $d_{ii} = d(X_i, X_i) = 0 \quad \forall i$ e che $d(X_i, X_j) = d(X_j, X_i)$ implica $d_{ij} = d_{ji}, \quad \forall i \neq j = 1, 2, ..., n$. Questo significa che le distanze coinvolte nel calcolo di s_d sono n^2, delle quali n sono sempre nulle mentre in generale $\frac{n^2-n}{2}$ sono diverse e non negative. Quindi $\bar{s}_d$ non è altro che la media aritmetica delle distanze non negative associate all'insieme X o equivalentemente all'insieme I. La matrice D diventa quindi utile per rappresentare le distanze tra gli elementi dell'insieme I.

1.1.3 Cluster analysis gerarchica

Anche se non sono le uniche, molte delle procedure di cluster sono gerarchiche, le quali implicitamente si basano sul concetto di distanza tra un oggetti e cluster.
Lo schema della Cluster analysis gerarchica consiste nel partire considerando gli n individui di I come un insieme di cluster $I_1, I_2, ..., I_n$, selezionare i due cluster di distanza più piccola, ad esempio I_i e I_j e unirli in un unico cluster. Il nuovo insieme di $n-1$ cluster sarà

$$I_1, I_2, ..., I_i, I_j, ..., I_n.$$

Si ripete il processo tante volte quante sono necessarie per ottenere un unico cluster di n individui, che non è altro che l'insieme originale I.

Verrà ora utilizzata la distanza euclidea al quadrato d_{ij}^2 come misura di distanza, cioè la matrice D sarà composta dalle distanze tra gli elementi al quadrato. Così, tenendo conto anche delle osservazioni precedenti si avrà:

$$D = d_{ij}^2 = \begin{bmatrix} 0 & d_{12}^2 & d_{13}^2 & \cdots & d_{1n}^2 \\ & 0 & d_{23}^2 & \cdots & d_{2n}^2 \\ & & 0 & \cdots & d_{3n}^2 \\ & & & \ddots & \vdots \\ & & & & 0 \end{bmatrix}.$$

Si supponga che gli individui I_i e I_j siano quelli più vicini, cioè tali che $d_{ij}^2 = min d_{ij}^2, i \neq j$, e che si decida di unirli in un unico cluster I_i, I_j. La matrice precedente diminuirà di dimensione, da $n \times n$ a $(n-1) \times (n-1)$ diventando:

$$D = \begin{bmatrix} 0 & d_{ij1}^2 & d_{ij2}^2 & d_{ij3}^2 & \cdots & d_{ijn}^2 \\ & 0 & d_{12}^2 & d_{13}^2 & \cdots & d_{1n}^2 \\ & & 0 & d_{23}^2 & \cdots & d_{2n}^2 \\ & & & 0 & \cdots & d_{3n}^2 \\ & & & & \ddots & \vdots \\ & & & & & 0 \end{bmatrix},$$

in cui $n-2$ righe sono rimaste le stesse, mentre una riga, la prima, è stata ricalcolata.
Nel ripetere il processo fino ad arrivare all'unico cluster finale, si ricalcola continuamente una riga della matrice con le distanze tra i cluster via via costituiti.

Dendrogrammi

Un aspetto importante della Cluster analysis gerarchica è la rappresentazione dei dati ottenuti e contenuti nelle matrici di distanza.
Il più comune metodo utilizzato per rappresentare una matrice di distanze è costituito dai *dendrogrammi* o *diagrammi ad albero*. In questi diagrammi tutti gli individui analizzati sono elencati verticalmente sulla sinistra mentre i risultati sono indicati verso destra. Il livello distanza è indicato orizzontalmente in alto.
Dati n oggetti ci sono molti possibili dendrogrammi che possano essere formati dai procedure gerarchiche di clustering; data però una determinata matrice

di distanze, ad essa corrisponde un solo diagramma ad albero.

Viene ora presentato un esempio di dendrogramma, rappresentazione di una Cluster analysis con sei elementi di partenza ($n = 6$). Gli oggetti 1 e 3 sono i più vicini, così sono raggruppati insieme al livello di distanza 1; seguono 5 e 4 al livello 2. Si noti che a questo punto del procedimento ci sono quattro cluster: $(1, 3), (6), (5, 4), (2)$. Nel terzo passaggio il procedimento forma il cluster $(1, 3, 6)$ e al quarto $(5, 4, 2)$, rispettivamente al livello di distanza 3 e 4. Infine tutti gli oggetti sono combinati insieme per formare un unico cluster con tutti gli oggetti al livello più grande, il quinto.

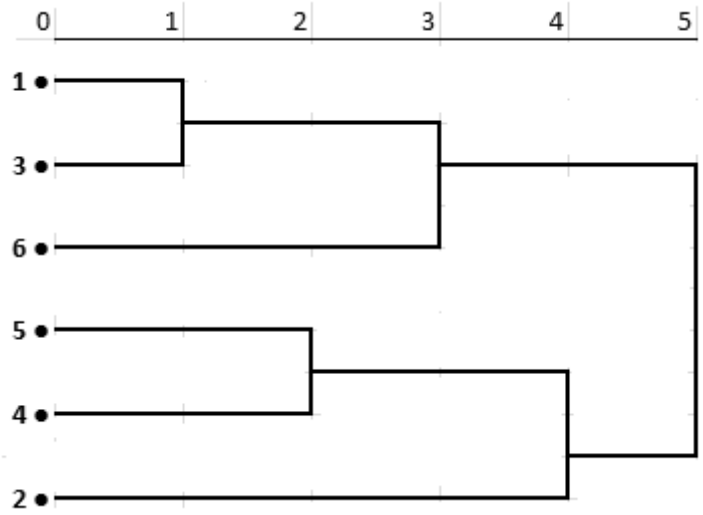


Figura 1.1: Esempio di dendrogramma

1.2 Analisi discriminante

Sia X un insieme di k campioni, suddivisi in p sottopopolazioni

$$X_1, X_2, ..., X_p$$

; l'analisi discriminante permette di assegnare una generica osservazione x ad una delle sottopopolazioni X_i.

Uno tra i primi studiosi a parlare di analisi discriminante multivariata fu R. A. Fisher nel 1936. Egli in quell'anno pubblicò in *The use of multiple measurements in taxonomic problems* un metodo astratto per la suddivisione di individui in gruppi, che non fosse legato ad una particolare scienza. Egli

riuscì con questo metodo ad attribuire alcuni reperti fossili alla categoria di umanoidi o di primati, in base a diverse misurazioni effettuate su di essi. Il suo obiettivo era quello di individuare la sottopopolazione di appartenenza di un'osservazione multidimensionale in base alla conoscenza campionaria del comportamento delle diverse sottopopolazioni, ma facendo anche in modo che non si fosse vincolati ad un certo ambito.

Problemi di questo genere sono tutt'ora di notevole interesse, sia da un punto di vista pratico che teorico. Se infatti si trattasse di osservazioni cliniche su un insieme di pazienti, da suddividere in due sottopopolazioni, quella composta di persone affette da una certa malattia e quelle sane, allora sarebbe utile disporre di una procedura automatica che, a partire dalle misurazioni, stabilisca se ogni nuovo individuo sia malato o no. D'altra parte, sarebbe anche interessante determinare la funzione delle misurazioni che meglio permetta di discriminare tra i vari gruppi.

L'assegnazione della suddetta osservazione x viene effettuata mediante una combinazione lineare $W = a'X$ delle k componenti della variabile X rilevata, in modo da massimizzare la discriminazione tra i p campioni. Il criterio che definisce la trasformazione, cioè il vettore di dimensione k di costanti a, consiste pertanto nel pretendere che sia massima la differenza tra le medie di W nei p campioni.
Da un punto di vista geometrico l'analisi discriminante consiste nel rappresentare i p campioni in uno spazio euclideo di dimensione $t < k$, tale da evidenziare opportunamente le distanze tra i campioni.

È necessario introdurre delle notazioni per poter spiegare il procedimento dell'analisi discriminante.
Siano:

- la matrice $n_j \times k$ del j-esimo campione:

$$X_j = \begin{bmatrix} x_{11_j} & \cdots & x_{1k_j} \\ \vdots & \ddots & \vdots \\ x_{n_j1_j} & \cdots & x_{n_jk_j} \end{bmatrix} = [x_{ihj}]\,;$$

- il j-esimo vettore k-dimensionale delle medie campionarie:

$$\bar{X_j} = \frac{1}{n_j} X_j' u_{n_j} = (\bar{X}_1, \bar{X}_2, ..., \bar{X_{kj}})';$$

- la j-esima matrice $k \times k$ delle varianze e covarianze campionarie:

$$S_j = \frac{1}{n_j}(X_j - u_{n_j}\bar{X}_j')'(X_j - u_{n_j}\bar{X}_j') = [S_{hlj}]\,.$$

Trasformando con il vettore a la matrice $n_j \times k$ del campione j-esimo, si ottiene per $j = 1, ..., p$ il vettore n_j-dimensionale

$$W_j = X_j a,$$

con:

- media campionaria: $\bar{W}_j' = a'\bar{X}_j'$,
- varianza campionaria: $S^2_{W_j} = a'S_j a$.

Posto $n = \sum_{j=1}^{p} n_j$ si indicando:

- il vettore k-dimensionale delle medie campionarie complessive con

$$\bar{X} = \frac{1}{n}X'u_n = (\bar{X}_1', ..., \bar{X}_k')';$$

- la matrice $k \times k$ delle varianze e covarianze campionarie complessive con

$$S = \frac{1}{n}(X - u_n\bar{X}')'(X - u_n\bar{X}') = [S_{hl}]\,.$$

La matrice S si può scomporre nel modo seguente: $S = S_w + S_b$, in cui:

- S_w indica la variaza *within*, cioè la matrice delle varianze e covarianze *all'interno* dei p campioni,
- S_b indica la varianza *between*, cioè la matrice delle varianze e covarianze *tra* i p campioni.

Analogamente trasformando con il vettore a la matrice X $n \times k$ di tutte le osservazioni disponibili si ottiene il vettore n-dimensionale $W = Xa$ con:

- media $\bar{W} = a'\bar{X}$
- varianza $S'_{W_a} = a'S_w a + a'S_b a$.

Si vuole ora definire W in modo tale da rendere massime le differenze tra le medie campionare $\bar{W}_1, ..., \bar{W}_p$. Ciò implica la massimizzazione di S_b di W rispetto ad a, ma quanto maggiori sono gli elementi del vettore a in valore assoluto , tanto più elevato è il valore della forma quadratica. Quindi affinché il problema sia ben definito, è necessario considerare un vincolo sulla dimensione di a dato dall'espressione $a'Sa = 1$, vincolo che corrisponde a pretendere che W abbia varianza unitaria.
Il problema da risolvere pertanto è di massimo vincolato, per il quale si può ricorrere al metodo dei moltiplicatori di Lagrange che si riduce alla risoluzione del sistema:

$$\begin{cases} \lambda = a'S_b a \\ a'Sa = 1 \end{cases}$$

in cui λ è il moltiplicatore di Lagrange ma anche uno degli autovalori di $S^{-1}S_b$ e a l'autovettore ad esso associato.
Affinché il valore di $a'S_b a$ sia massimo, tra gli autovalori λ è necessario prendere quello con valore massimo. Sia questo indicato con λ_1. Allora, la variaile che si definisce mediante l'autovettore a_1 ad esso associato sarà

$$W_1 = a_1'X$$

ed è la combinazione lineare delle componenti della variabile k-dimensionale di partenza che separa maggiormente i p campioni. La variabile W_1 è detta *prima funzione discriminante lineare*; l'autovalore λ_1 è detto *potere discriminante* di W_1 e ne misura la capacità di separare le medie dei p campioni (corrisponde quindi alla varianza between di W_1).
Per definire la *seconda funzione discriminante lineare* W_2 è necessario che, oltre ad essere soddisfatti i due vincoli precedenti, essa sia incorrelata con W_1, il che si traduce nell'espressione $a_1'Sa_2 = 0$. Anche questo problema è risolvibile mediante il metodo dei moltiplicatori di Lagrange, con il quale si determina il secondo maggiore autovalore della matrice $S^{-1}S_b$, λ_2, e l'autovettore corrispondente a_2

1.3 Analisi della varianza e Test di Bonferroni

1.3.1 Analisi della varianza (ANOVA)

L'*Analisi della varianza* (brevemente detta ANOVA, chiamandosi *Analysis of Variance* in inglese) è un insieme di tecniche che rientrano nella statistica inferenziale. Essa si applica a due o più gruppi di dati e, mediante lo studio della variabilità all'interno dei gruppi e tra i gruppi, consente

di trarre informazioni sulla differenza delle medie tra gli stessi gruppi. Se i gruppi sono solo due l'Analisi della varianza coincide con la *t di Student*.
L'ipotesi alla base dell'Analisi della varianza è che, dato un certo numero di gruppi in esame, sia possibile scomporre la varianza totale in due componenti: varianza interna ai gruppi (anche detta *varianza within*) e varianza tra i gruppi (*varianza between*). La ragione della distinzione tra le due varianze è la convinzione che determinati fenomeni trovino spiegazione in caratteristiche proprie del gruppo di appartenenza. Vale a dire, se la variabilità interna ai gruppi è relativamente elevata rispetto alla variabilità tra i gruppi, allora è probabile che la differenza tra questi gruppi sia soltanto il risultato della variabilità interna.
Il procedimento è il seguente.
Siano:

- $SQQ_b = \sum_{i=1}^{k} n_i(m_i - m)^2$ la variabilità fra i gruppi,
- $SQQ_w = \sum_{i=1}^{k} \sum_{j=1}^{n_j}(x_{ij} - m_j)^2$ la variabilità interna ai gruppi,
- $SQQ_{tot} = SQQ_b + SQQ_w = \sum_{i=1}^{k} \sum_{j=1}^{n_j}(x_{ij} - m)^2$ la variabilità totale,

dove k è il numero dei gruppi, n_j e n rispettivamente la numerosità del j-esimo e la numerosità complessiva delle osservazioni, m_j e m la media del j-esimo gruppo e la media generale, infine x_{ij} l'i-esimo elemento del j-esimo gruppo.
Dividendo le somme dei quadrati precedenti per i relativi gradi di libertà si ottengono le varianze descritte sopra; in particolare i gradi di libertà di SQQ_b sono $(k-1)$ mentre quelli di SQQ_w sono $(n-k)$.
Per confrontare le due varianze si calcola il rapporto

$$F = \frac{\frac{SQQ_b}{k-1}}{\frac{SQQ_w}{n-k}} = \frac{SQQ_b(n-k)}{SQQ_w(k-1)}$$

e lo si confronta con i valori della F di Fisher.
Se F risulta significativo, quindi SQQ_b abbastanza grande, si può presumere l'esistenza di differenze fra le medie dei gruppi.
Viceversa, se F non risulta significativo, si può concludere che le differenze tra le medie dei gruppi possono avere origine casuale.
Si noti che per significativo si intende maggiore dei limiti previsti dalle apposite tabelle per i gradi di libertà di interesse.

1.3.2 Test di Bonferroni

Il *test di Bonferroni* permette di integrare i risultati dell'analisi della varianza; infatti con esso si calcolano tutte le differenze tra le medie di ogni gruppo con quelle degli altri gruppi, al fine di individuare quanti e quali gruppi rendono significativo il test F tra le varianze between e within. Il test di Bonferroni consiste nel confronto tra il valore t di Bonferroni, da ricercare sulle tavole t di Student, con il valore

$$t_{Bonferroni} = \frac{m_i - m_j}{\sqrt{S_e^2(\frac{1}{n_1} + \frac{1}{n_j})}}$$

dove S_e è la varianza entro i gruppi.

Questo test permette quindi di contrastare il problema dei confronti multipli, nonostante sia adatto per un numero non troppo grande di test simultanei. Si rivela utile perché, quando vengono svolti numerosi test, alla fine potrebbe accadere che si termini con un risultato che mostra significatività statistica anche se non ne ha. Se un particolare test riporta risultati corretti il 99% delle volte, facendo 100 volte, potrebbe riportare qualche falso risultato. Il test di Bonferroni abbassa il livello di confidenza di ogni confronto per ottenere il livello di confidenza complessivo desiderato.
Alla base di questo metodo c'è la disuguaglianza di Bonferroni. Dati due insiemi A e B:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B),$$

la quale, generalizzando, implica:

$$P(\cup_{i=1}^{\infty} A_i) =\leq \sum_{i=1}^{\infty} P(A_i).$$

Considerando il complementare di A_i, cioè A_i^c, e sostituendolo nella precedente disuguaglianza, per un numero finito di insiemi si avrà:

$$P(\cup_{i=1}^{n} A_i^c) \leq \sum_{i=1}^{n} P(A_i^c) = P(\cap_{i=1}^{n} A_i)^c \leq \sum_{i=1}^{n}(1 - P(A_i)) \tag{1.1}$$

$$= 1 - P(\cap_{i=1}^{n} A_i) \leq n - \sum_{i=1}^{n} P(A_i) \tag{1.2}$$

$$= P(\cap_{i=1}^{n} A_i) \geq \sum_{i=1}^{n} P(A_i) - (n-1), \tag{1.3}$$

che è proprio la disuguaglianza di Bonferroni.
Si supponga, per esempio, di voler determinare una regione di confidenza di livello di confidenza globale pari a $1-\alpha$. A tale scopo, considerando h confronti ciascuno con livello γ, dove γ è tale che:

$$1-\alpha = \sum_{i=1}^{h} \gamma - (h-1).$$

Per garantire un livello globale di confidenza uguale a $1-\alpha$, ogni confronto deve avere livello di confidenza uguale a

$$\gamma = 1 - \frac{\alpha}{h}.$$

1.4 La simulazione

In molti ambiti scientifici e tecnologici, la riproduzione di situazioni che si debbono analizzare non è sempre possibile o è molto difficile, per diverse ragioni. Un potente strumento sperimentale di analisi, che permette di ovviare a questo problema, è stato fornito dallo sviluppo dell'informatica e dei calcolatori: la simulazione dei dati. Con la simulazione si traduce la realtà in un modello concettuale o matematico, definito come insieme di processi che hanno luogo nel sistema valutato e che consente di comprendere le logiche del funzionamento del sistema stesso, di valutare e prevedere lo svolgersi di eventi, avendo solo imposto alcune condizioni. I vantaggi di questo laboratorio virtuale che viene così costituito sono molteplici e dipendono dall'ambito scientifico in cui si opera.
In quello che interessa la trattazione, la simulazione ha consentito di disporre, in poco tempo, di dati relativi a popolazioni simili a quelle esistenti o differenti in qualche caratteristica, così da studiarne le conseguenze in termini matematici, con i metodi descritti in questo capitolo.

1.4.1 Il simulatore *msms* e la distribuzione binomiale

Per la simulazione di popolazioni e dati demografici è stato utilizzato un software chiamato *msms*, al quale si forniscono dei precisi comandi che verranno spiegati in seguito.
In questa sede si vuole solo approfondire un concetto matematico alla base del simulatore: il numero di individui in ogni successiva generazione è una variabile casuale binomiale, con un parametro dipendente dalla precedente generazione.

Siano effettuate n prove, indipendenti tra loro, ognuna delle quali con probabilità di successo pari a p. Se X rappresenta il numero di successi verificatisi tra le n prove, allora si dice che X è una variabile casuale binomiale di parametri (n,p) ($\mathcal{B}(n,p)$). La distribuzione di probabilità è data da:

$$P_i \equiv \ P\{X=i\} = \binom{n}{i} p^i (1-p)^{n-i}, \quad i = 0,1,...,n, \tag{1.4}$$

in cui ogni successione con i successi e $(n-k)$ insuccessi ha probabilità $p^i(1-p)^{n-i}$, e il numero di tali successioni, pari al numero di combinazioni in cui possono essere disposti gli i successi negli n tentativi, è dato dal coefficiente binomiale

$$\binom{n}{i} = \frac{n!}{i!(n-i)!}.$$

Il fatto che una variabile X casuale binomiale (n,p) rappresenti il numero di successi in n prove indipendenti, ognuna con probabilità di successo p, si rappresenta con:

$$X = \sum_{i=1}^{n} X_i \tag{1.5}$$

dove:

$$X_1 = \begin{cases} 1 & \textit{se l'i-esima prova ha successo} \\ 0 & \textit{altrimenti} \end{cases}.$$

Inoltre valgono:

$$E[X_i] = P\{X_i = 1\} = p$$

$$Var(X_i) = E[X_i^2] - E([X_i])^2 = p - p^2 = p(1-p),$$

così da (1.5), per una variabile casuale binomiale (n,p), segue:

$$E[X] = \sum_{i=1}^{n} E[X_i] = np$$

$$Var(X) = \sum_{i=1}^{n} = np(1-p)$$

essendo gli X_i indipendenti.

Infine, vale la seguente formula ricorsiva, la quale permette di esprimere la probabilità p_{i+1} in termini di p_i:

$$p_{i+1} = \frac{n!}{(n-i-1)!(i+1)!} p^{i+1}(1-p)^{n-i-1} \tag{1.6}$$

$$= \frac{n!(n-i)}{(n-i)!i!(i+1)!} p^i (1-p)^{n-i} \frac{p}{1-p} \tag{1.7}$$

$$= \frac{n-i}{i+1} \frac{p}{1-p} p_i \tag{1.8}$$

Caso particolare della distribuzione binomiale è la distribuzione di Bernoulli, nella quale $n = 1$: $\mathcal{B}(1, p)$.

1.5 L'interpolazione

In analisi numerica, si parla di interpolazione quando, a partire da un insieme di punti dati, interpretabili come punti di un piano (x, y), ci si propone di costruire una funzione che sia in grado di descrivere la relazione che intercorre fra l'insieme dei valori x e l'insieme dei valori y.
Vale a dire: siano dati $n + 1$ numeri reali distinti detti nodi x_i; per ognuno di questi sia dato un secondo numero y_i. Si vuole individuare una funzione f tale che

$$f(x_i) = y_i \quad per\ i = 0, \ldots, n.$$

Una coppia (x_i, y_i) costituisce un punto dato ed f viene detta funzione interpolante per tali punti e appartiene ad un certo spazio di funzioni definite, il quale determina anche il procedimento da seguire.

Spesso, nelle attività scientifiche e tecnologiche e in tutti negli studi quantitativi di qualsiasi fenomeno, accade di disporre di un certo numero di punti del piano ottenuti per campionamento o in altro modo, e di ritenere opportuno individuare una funzione che passi per tutti i punti dati o almeno nelle loro vicinanze. Si parla allora di curve fitting. Le curve approssimanti possono essere usate come aiuto per visualizzare i dati, per rappresentare i valori di una funzione dove non sono disponibili i dati, e per riassumere le relazioni tra due o più variabili.

Le funzioni interpolanti, e i metodi per trovarle, sono di diversi tipi.
Ad esempio, si può scegliere di interpolare gli $n + 1$ punti dati mediante dei

polinomi (interpolazione polinomiale) di grado inferiore o uguale ad n:

$$f(x) = p(x) = a_0 + a_1 x + ... + a_n x^n.$$

Il problema dunque consiste nel determinare i coefficienti a_i tali che

$$p(x_i) = \sum_{j=0}^{n} a_j x_i^j = f(x_i), \quad i = 0, ...n.$$

L'interpolazione razionale, invece, è simile a quella polinomiale ma utilizza delle funzioni razionali della forma

$$R(x) = \frac{P(x)}{Q(x)},$$

in cui $P(x)$ e $Q(x)$ sono due polinomi.
Ancora, l'interpolazione trigonometrica utilizza invece delle funzioni interpolanti della forma

$$f(x) = a_0 + a_1(cosx + senx) + ...$$

Nel corso della trattazione sono state provate due diverse funzioni interpolanti e la scelta della migliore è avvenuta confrontando le curve mediante un indice: il coefficiente di determinazione R^2, il quale fornisce indicazioni riguardanti la bontà di adattamento di un modello statistico ai dati.
Considerate n osservazioni, si calcola nel modo seguente:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS},$$

dove

- $ESS = \sum_{i=0}^{n-1} (\hat{y}_1 - \bar{y})^2$, *Explained Sum of Squares*, è la devianza spiegata dal modello,
- $TSS = \sum_{i=0}^{n-1} (y_i - \bar{y})^2$, *Total Sum of Squares*, è la devianza totale,
- $RSS = \sum_{i=0}^{n-1} (y_i - \hat{y}_i)^2$, *Residual Sum of Squares*, è la devianza residua,

con:

- y_i dati osservati,

- $\bar{y}$ media delle osservazioni,
- $\hat{y}_i$ dati stimati dal modello statistico utilizzato.

Esso può variare tra 0 e 1: un valore uguale a 1 indica un adattamento perfetto del modello ai dati, al contrario un valore uguale a 0 indica che il modello utilizzato non è utile a spiegare i dati.

Capitolo 2

Introduzione al tema antropologico

2.1 La genetica evolutiva

La genetica evolutiva si basa sul principio che l'*archivio genetico* della vita è contenuto nei genomi delle specie viventi. La decodifica di questo archivio rivela la storia evolutiva e le relazioni filogenetiche delle specie studiate. La capacità di leggere le informazioni genetiche si è sviluppata enormemente negli ultimi anni.
Le prove genetiche provengono da due principali fonti:

1. dai genomi di individui viventi, necessariamente tramandati dagli antenati;

2. dal DNA antico di resti organici ben conservati, che può figurare o no nei discendenti viventi.

Grazie alla disponibilità di genomi umani e non umani e allo sviluppo di nuove tecnologie, le quali permettono di analizzare la maggior parte delle mutazioni genetiche, tutte le scienze collegate alla genetica evolutiva stanno conoscendo un enorme progresso e si sta dunque vivendo un periodo di innovazione che fornisce potenti strumenti per analizzare insiemi di dati di ineguagliabile qualità e quantità; inoltre la ricchezza stessa di dati catalizza lo sviluppo di nuovi metodi interpretativi dei fenomeni evolutivi.

Dalla suddetta ricchezza di dati sulle variazioni genetiche tra umani, e tra umani e altri primati di cui si dispone oggi, è possibile rispondere a molte delle domande fondamentali riguardo all'origine umana. Ad esempio si può sapere quanto sono geneticamente diversi gli umani, quando gli umani

si sono divisi dai Neandertaliani o quando e dove l'uomo ha avuto origine. Si può rispondere a domande relative a individui e gruppi, o quale sia negli afroamericani la proporzione di geni provenienti dall'Africa. Ancora, più recentemente, è stato possibile rispondere a domande sui fenotipi, cioè quali cambiamenti genetici ci hanno resi umani o ci hanno permesso di adattarci a differenti condizioni climatiche.

Ci sono però anche domande sull'origine dell'uomo a cui non si può rispondere e non si potrà neanche in futuro. Quello che vorremmo conoscere e quello che potremmo conoscere sono due cose diverse a causa di limiti tecnici ed etici. Non si potrà mai sapere, ad esempio, se due specie australopitecine contemporanee fossero interfeconde o quali combinazioni di mutazioni sarebbero sufficienti per rendere gli scimpanzè capaci di parlare. Si noti però anche che non tutti i limiti sono statici: la divergenza genetica tra Neandertaliani e uomini moderni era, per esempio, inconoscibile fino a prima dell'avvento dei metodi di sequenziamento del DNA, mentre oggi si può misurare tale divergenza su tutto il genoma. In conclusione, si consideri che un importante aspetto degli studi nella genetica evolutiva umana è il porsi domande chiare e alle quali si possa dare una risposta.

2.2 Il genoma umano e la sua variazione

Per comprendere il soggetto della genetica evolutiva è necessaria una conoscenza di base del genoma umano.

La maggior parte degli organismi viventi utilizza l'acido desossiribonucleico, il DNA, come materiale genetico. Il DNA è una macromolecola che svolge due funzioni biologiche fondamentali: contiene e trasmette le informazioni necessarie al corretto funzionamento della cellula e dell'intero organismo, tanto che talvolta viene chiamato "molecola della vita".
Il DNA è un polimero, cioè una molecola costituita da monomeri, chiamati nucleotidi, che sono unità che si ripetono un numero elevato di volte. I monomeri sono a loro volta costituiti da tre parti chimicamente distinte: un gruppo fosfato, uno zucchero, cioè il desossiribosio, e una base azotata. Le basi sono quattro: adenina, guanina, citosima, timina (rispettivamente abbreviate con A, G, C, T); è l'ordine nella disposizione sequenziale dei nucleotidi che costituisce l'informazione genetica.

Il DNA non è l'unico acido nucleico presente nelle cellule; ci sono infatti diversi tipi di acidi ribonucleici, RNA, che giocano ruoli fondamentali, spe-

cialmente nella produzione di proteine. L'RNA e il DNA differiscono per il tipo di molecola di zucchero che contengono (ribosio anziché desossiribosio) e in una base, la timina che è sostituita dall'uracile (U).

La struttura del DNA che Watson e Crick descrissero nel 1953 è costituita da due filamenti polinucleotidi appaiati tra loro. I due filamenti sono tenuti assieme dai legami ad idrogeno che si instaurano tra le basi azotate. L'accoppiamento delle basi avviene sempre e solo in un unico modo a causa della loro struttura chimica e precisamente: G-C, T-A. Questo sistema di appaiamento costituisce la regola della complementarità delle basi.
I due filamenti che costituiscono la molecola sono inoltre antiparalleli, ovvero sono orientati in direzioni opposte.
La molecola di DNA presenta una forma a doppia elica dato che i due filamenti che la costituiscono si avvolgono intorno all'asse di allungamento della molecola stessa.
Negli eucarioti, il DNA è solitamente presente all'interno dei cromosomi. La somma di tutti i cromosomi di una cellula ne costituisce il genoma; il genoma umano conta circa 3 miliardi di paia di basi contenute in 46 cromosomi. La disposizione finale nei cromosomi segue precise regole gerarchiche di impacchettamento. Nelle cellule, infatti, il doppio filamento di DNA non può essere disposto a casaccio poiché la lunghezza dei filamenti di DNA è solitamente molto elevata e creerebbe seri problemi alla cellula ospite.
Nel genoma, l'informazione è conservata in sequenze di DNA chiamate geni. La trasmissione dell'informazione contenuta nei geni è garantita dalla presenza di sequenze di basi azotate complementari. Infatti, durante la trascrizione, l'informazione può essere facilmente copiata in un filamento complementare di RNA. Solitamente, tale copia di RNA è utilizzata per sintetizzare una proteina, attraverso un processo definito traduzione. In alternativa, una cellula può semplicemente duplicare l'informazione genetica attraverso un processo definito replicazione del DNA.

Un filamento di DNA solitamente non interagisce con altri segmenti di DNA e, nelle cellule umane, i differenti cromosomi occupano addirittura regioni separate del nucleo. Tale separazione fisica è fondamentale per permettere al DNA di essere un archivio stabile e sicuro dell'informazione genetica. L'interazione tra diversi segmenti di DNA è invece possibile e frequente attraverso il fenomeno del crossing-over, che permette la ricombinazione genetica attraverso la rottura di due eliche, lo scambio di segmenti tra di esse ed il ricongiungimento finale.
La ricombinazione permette ai cromosomi di scambiare informazioni genetiche e produrre nuove combinazioni di geni, con il risultato di aumentare

l'efficienza della selezione naturale e di facilitare l'evoluzione di nuove proteine. La ricombinazione genetica può anche essere coinvolta nella riparazione del DNA, in particolare come risposta cellulare in seguito a rotture a doppio filamento.

Gli oltre 7 miliardi di persone viventi oggi contengono 14 miliardi di copie di genoma tutte differenti, con le rare eccezioni dei gemelli identici. Le differenze tra i genomi sono da polimorfismi a singolo nucleotide (SNPs) fino a variazioni strutturali che coinvolgono milioni di paia basi della sequenza del DNA. Molte differenze non hanno effetti riscontrabili sulle persone che le presentano, alcune risultano senza effetti deleteri, e altre causano, predispongono o proteggono da gravi malattie.

La più semplice e piccola differenza tra due analoghe sequenze di DNA è la sostituzione di una base con un'altra. Lo scambio di A con G o di C con T e viceversa viene chiamato transizione, mentre lo scambio di C o T con G o A e viceversa è detto transversione. Transizione e transversione sono esempi di SNPs; oltre a questi, sono SNPs anche le inserzioni e le cancellazioni (indel) di singole basi.

2.3 La genetica delle popolazioni umane

La scienza che si occupa di come tutti i principi genetici si applichino ad intere popolazioni di organismi interfecondi è la genetica delle popolazioni umane, una branca della genetica. Essa studia il fenomeno del polimorfismo genico, che è alla base della biodiversità genetica in quasi tutte le specie, e per quanto riguarda la specie umana analizza le popolazioni contemporanee su diversa scala, le affinità genetiche fra gruppi, e i motivi delle differenze nell'incidenza di malattie ereditarie e di altri caratteri.
Occupandosi delle proprietà genetiche e della loro trasmissione in gruppi di individui, la genetica delle popolazioni (umane) è la disciplina che per eccellenza osserva e analizza i cambiamenti elementari alla base del processo di evoluzione. L'insieme dei genomi di individui fra loro interfecondi a diversi livelli di raggruppamento (comunità locali, popolazioni regionali, continentali, intera specie) costituisce un pool genetico all'interno del quale è presente una notevole quota di variabilità genetica, a causa dell'esistenza di forme varianti (alleli) in pressoché ogni tratto di DNA. In questo quadro l'evoluzione viene considerata un processo interamente variazionale in base al quale non si verifica una graduale trasformazione del singolo da una condizione a un'altra ma si modifica la composizione del gruppo a cui la sua discendenza appartiene,

e di conseguenza si modifica la composizione del pool genetico attraverso le generazioni. Sono quindi i fenomeni a livello di popolazione che costituiscono i cambiamenti elementari alla base della diversificazione fra gruppi ed è la progressiva diversificazione che, insieme ad altri fattori di tipo geografico, comportamentale, sociale ecc., può condurre a un isolamento riproduttivo di grado variabile.

Comprendendo i meccanismi con cui i processi evolutivi agiscono si possono produrre modelli matematici approssimanti la realtà. Tali modelli sono necessari per comprendere la sottile interazione tra i processi e permettono di inferire sui processi passati a partire dalla diversità moderna; si possono stimare infatti i tassi di crescita delle popolazioni, le età degli alleli, i tassi di migrazione tra due popolazioni e fare diverse ipotesi sul passato.
Un esempio di semplice modello nella genetica delle popolazioni è l'equilibrio di Hardy-Weinberg, il quale postula che all'interno di una popolazione, da una generazione all'altra, c'è equilibrio tra le frequenze alleliche e le frequenze genotipiche, cioè queste non cambiano con il passare del tempo a meno che non intervengano fattori specifici atti a disturbare l'equilibrio stesso (migrazioni, selezione, mutazione, ...).

Esistono diversi fenomeni che hanno vari effetti sulla variabilità genetica e talvolta agiscono anche insieme; essi sono le mutazioni, la ricombinazione, la deriva genetica, il flusso genico, la selezione naturale.

Mutazione e ricombinazioni incrementano la diversità umana generando rispettivamente nuovi alleli e nuovi aplotipi. Per mutazioni si intendono tutte le modifiche stabili o ereditabili nel materiale genetico dovute ad agenti esterni o al caso, ma non alla ricombinazione genetica. Non sono un fenomeno molto frequente in natura, a differenza delle ricombinazioni, le quali rappresentano la sorgente più importante di variazione genetica.
La deriva genetica consiste in un fenomeno casuale di oscillazione delle frequenze alleliche, correlato al numero di individui che costituiscono una popolazione. Quando eventi perturbatori come terremoti, alluvioni, incendi distruggono gran parte di una popolazione facendo vittime in maniera non selettiva (effetto "collo di bottiglia") oppure quando pochi individui colonizzano un nuovo ambiente per la specie, come un lago o un'isola ("effetto fondatore"), probabilmente la nuova e più piccola popolazione che ne deriverà presenterà caratteristiche totalmente diverse da quelle della popolazione di origine. In questo caso l'effetto della deriva genetica perdura fino a che la popolazione non sarà più grande.
Con il flusso genico si diffondono geni tra popolazioni per migrazione di indi-

vidui in età riproduttiva; di conseguenza nella popolazione ricevente possono cambiare le frequenze alleliche e generarsi nuovi geni. Globalmente il flusso genico aumenta il polimorfismo di una popolazione e, allo stesso tempo, riduce le differenze genetiche medie tra le popolazioni.
La selezione naturale, infine, è il meccanismo con cui avviene l'evoluzione delle specie e secondo cui, nell'ambito della diversità genetica delle popolazioni, si ha un progressivo e cumulativo aumento della frequenza degli individui con caratteristiche ottimali per l'habitat, e che avviene per il diverso successo riproduttivo. La tendenza della selezione naturale è di ridurre la variabilità genetica.

2.4 Out of Africa: le espansioni demografiche dell'uomo

La storia dell'evoluzione umana è fortemente legata alle diverse espansioni fuori dall'Africa, avvenute quando le condizioni lo permisero, ma ad una di queste in particolar modo. È quella che ebbe luogo circa 50-70 mila anni fa, la quale ha portato alla definitiva estinzione dei primi ominidi, anche se non senza incroci, e successivamente all'occupazione dell'intero globo. Tale espansione ha marcato l'inizio di una transizione dalla demografia scimmiesca ad una moderna demografia e i modelli geografici stabiliti a quell'epoca persistono ancora oggi e forniscono lo sfondo per la comprensione di gran parte della genetica umana.

Confrontando geneticamente le popolazioni africane e non africane si perviene a due semplici conclusioni: le popolazioni non africane contengono un piccolo sottoinsieme della diversità africana e questo sottoinsieme è condiviso da tutte le popolazioni fuori dall'Africa. Nonostante si tratti di eccessive generalizzazioni (poiché non considerano le recenti migrazioni, come quelle indietro verso il nord e l'est Africa, quelle legate alla tratta degli schiavi e alle colonizzazioni), forniscono la chiave per la comprensione delle numerose caratteristiche delle nostre variazioni genetiche.
Queste osservazioni suggeriscono quindi l'esistenza di un'unica popolazione africana dalla quale sono poi derivate tutte le popolazioni non africane.

Tra le teorie paleoantropologiche che tentano di descrivere come è avvenuta la popolazione del globo da parte dell'uomo, la più accreditata è la cosiddetta teoria "Out of Africa", secondo la quale in tre grandi ondate migratorie, tre diverse specie di Homo originarie dell'Africa si spinsero fuori dal

continente africano popolando gli altri e adattandosi alle condizioni climatiche trovate.
La prima migrazione fuori dall'Africa (Out of Africa I) iniziò circa 1,5 milioni di anni fa per opera dell'Homo erectus; si formarono così gruppi di Homo Erectus in Asia (ne sono esempi l'Uomo di Pechino e l'Uomo di Giava) e in Europa (come l'Homo antecessor e l'Uomo di Ceprano). Un ceppo di erectus rimase invece in Africa: l'Homo ergaster. Fu proprio questo ceppo che nei millenni successivi diede origine ad un uomo più evoluto, denominato Homo heidelbergensis (o Homo sapiens arcaico), che fece la sua comparsa 600 mila anni fa ed è il ceppo di origine dell'uomo moderno.
Questo sapiens arcaico tra i 500 mila e i 250 mila anni fa diede origine a dei movimenti migratori in Eurasia (Out of Africa II). In Europa ne sono una testimonianza le popolazioni pre-neandertaliane, le quali successivamente sfociarono nell'Uomo di Neanderthal, vissuto tra i 100 mila e i 30 mila anni fa, quando si estinse, ma che si spostò dall'Europa anche in Medio Oriente. Le popolazioni emigrate in Asia orientale diedero origine ad una esigua popolazione di Homo sapiens, l'uomo di Denisova, già in declino 125 mila anni fa e successivamente estintasi.
Il ceppo rimasto in Africa evolvette verso l'Homo sapiens sapiens che fece la sua comparsa in Africa del sud attorno a 130, forse 190, mila anni fa: è l'uomo moderno. Homo sapiens sapiens diede a sua volta luogo a diverse migrazioni fuori dall'Africa (Out of Africa III) raggiungendo il Medio Oriente 90 mila anni fa e l'Eurasia tra i 40 mila e i 10 mila anni fa. In Europa giunse tra i 30 e i 40 mila anni fa: l'Uomo di Cromagnon.
In Eurasia occidentale, l'Homo sapiens sapiens incontrò il Neanderthal, destinato ad estinguersi; in piccola parte si incrociò con esso, dando origine alle attuali popolazioni euroasiatiche e americane, le quali possiedono il $2,5\%$ di DNA neandertaliano. Continuando la sua migrazione in Australia e in Asia orientale, Homo sapiens sapiens incontrò, sostituendosi e in parte incrociandosi, anche i Denisoviani, dando origine alle attuali popolazioni della Melanesia, della Nuova Guinea e Australia, le quali hanno tra il $2 \div 5\%$ di geni neandertaliani e per il 5% geni denisoviani.
Il ceppo di Homo sapiens sapiens che stazionò in Africa diede invece origine alle attuali popolazioni africane, con lo 0% di DNA neandertaliano e lo 0% di DNA denisoviano.
Tutte le popolazioni dell'uomo moderno derivano quindi dall'Homo sapiens originatosi in Africa e da qui migrato in tutto il resto del mondo.

2.5 *Progetto 1000 genomi umani*

La teoria Out of Africa è supportata dalla genetica e in particolare dagli ultimi studi condotti nel *progetto 1000 genomi umani*. Tale progetto è stato avviato con lo scopo di catalogare le varianti genetiche umane, cioè le piccole differenze in specifiche regioni del genoma che rendono ogni individuo unico, diverso da tutti gli altri, in modo da tracciare la prima mappa completa dei geni collegati alle malattie, rare e meno rare, che affliggono il genere umano. Sono stati sequenziati per ora i genomi di 1092 persone provenienti da 20 popolazioni in Europa, Asia orientale, Africa sub-sahariana e le Americhe. In ultima analisi, si studieranno più di 2500 persone provenienti da 26 popolazioni.
Oltre che in medicina, biologia e altri ambiti scientifici, il progetto 1000 genomi ha fornito una banca dati con la quale si è potuto dimostrare che la diversità nelle popolazioni africane comprende anche la maggior parte di quella di tutte le altre popolazioni mondiali. Ciò vuol dire che le varianti che vengono identificate in popolazioni provenienti da tutto il resto del mondo sono quasi sempre presenti anche nelle popolazioni africane, mentre non è vero il contrario. Da un punto di vista evolutivo questo dato conferma l'origine africana della popolazione attuale di Homo sapiens.

L'elaborato parte proprio dall'analisi dei dati forniti dal progetto, nel loro complesso in un primo momento, per poi passare allo studio di alcuni di essi nel dettaglio; di questi se ne osservano misure di distanza e i possibili raggruppamenti mediante la Cluster Analysis e l'analisi discriminante. Conoscendo la storia delle popolazioni studiate inoltre è stato possibile osservare alcune loro caratteristiche *da un punto di vista matematico*, delle quali si parlerà più dettagliatamente in seguito.
La seconda parte della tesi invece è concentrata sull'analisi di dati relativi a popolazioni generate da un simulatore, volta a trovare dei metodi per risalire a delle informazioni sulle popolazioni reali.

Capitolo 3

Materiali e metodi

3.1 I dati del *progetto*

I dati da cui si è partiti sono tratti dalle pubblicazioni del *Progetto 1000 genomi umani* e si riferiscono alle popolazioni elencate di seguito:

- AMHARA - Etiopi
- ASW - AfroAmericani
- CEU - Europei
- CHB - Cinesi Nord
- CHS - Cinesi Sud
- CLM - Colombiani
- GIH - Indiani
- GUMUZ - Etiopi Ovest
- IBS - Spagnoli
- LWK - Kenyoti
- MXL - Messicani
- NATAM - Peruviani
- OROMO - Etiopi Est
- PEL - Peruviani
- PUR - Portoricani
- SOMALI - Somali
- TSI - Toscani
- WOLAYTA - Etiopi Sud
- YRI - Nigeriani

Tali dati si presentano come matrici, composte da 20 colonne: mentre la prima colonna è occupata dalle classi di frequenze relative, quindi da una suddivisione dell'intervallo $[0, 1]$, le altre 19 sono destinate ognuna ad una popolazione delle precedenti. Presa una popolazione di riferimento, nelle colonne si legge la frequenza relativa di SNPs che tale popolazione condivide con ognuna delle altre. Evidentemente sono dati che hanno senso se considerati a coppie di popolazioni.

Di seguito ne viene riportato un esempio: la popolazione di riferimento è quella afroamericana (ASW); lo si evince dal fatto che la colonna destinata agli ASW è occupata da tutti 1, poiché tutte le mutazioni presenti in tale popolazione sono condivise da se stessa. Per ragioni di spazio sono riportate solo le frequenze di 4 popolazioni (ASW, AMHARA, CEU, TSI).

BIN	ASW	AMHARA	CEU	TSI
0.01	1	0.185142027518453	0.174401168752533	0.183263059581336
0.02	1	0.307346470003742	0.21365336223917	0.229649645774457
0.031	1	0.406978002163722	0.252788250336033	0.274215650919582
0.041	1	0.4995020932522	0.298172903653625	0.323356652249891
0.051	1	0.573030876604921	0.347730988531882	0.376700147087262
0.061	1	0.640392865493048	0.398049084624851	0.429212257200389
0.071	1	0.698761319187404	0.450243730532749	0.482702464885073
0.082	1	0.749217565339038	0.495126695700566	0.530056723321914
0.092	1	0.794410063172542	0.543126003057773	0.581964984696517
0.102	1	0.832360013740982	0.587463335359247	0.622469809475993
0.112	1	0.85818328128576	0.620787284150222	0.656740714491373
0.122	1	0.882175324543734	0.657878418196335	0.692328892599633
0.133	1	0.900966194216933	0.690297964956184	0.724848329977574
0.143	1	0.922328126208337	0.728230608614956	0.758318188848504
0.153	1	0.934947408564081	0.752245636276495	0.782795565931263
0.163	1	0.947580956584101	0.783472168588294	0.810189220499807
0.173	1	0.957103175068272	0.804034176814618	0.83075678906717
0.184	1	0.965139366653177	0.828011179684338	0.852425637393768
0.194	1	0.974604763805362	0.845829133490729	0.871894089553047
0.204	1	0.978837866377465	0.858871543137647	0.882595108307885
0.214	1	0.983276323673293	0.873704009890102	0.896593830821622
0.224	1	0.9874453626009	0.889042210381903	0.909716683150943
0.235	1	0.990405897918088	0.900497595233737	0.92050125646957
0.245	1	0.991785612270512	0.911589297925196	0.929048117788607
0.255	1	0.993845783667728	0.922919597245872	0.939531354112682
0.265	1	0.995161862060376	0.931806206229222	0.94677090219297
0.276	1	0.996037159101092	0.940288550576006	0.95441737277588
0.286	1	0.997196756543497	0.946489197130309	0.956622403106409
0.296	1	0.997921245520392	0.952145339584033	0.961088314584845
0.306	1	0.998710207622331	0.956181301434751	0.967127415507185
0.316	1	0.999036001190122	0.959666765843499	0.969080630764653
0.327	1	0.999118079372856	0.967148456638902	0.974816266536012
0.337	1	0.999078977665208	0.971307595875867	0.978483894901118
0.347	1	0.999277466139436	0.972911548718488	0.980241983158393
0.357	1	0.999581301746377	0.976660948959332	0.98286038439201
0.367	1	0.999543858679125	0.978796339811462	0.984767644375639
0.378	1	0.99971622350237	0.978589063253781	0.986094951616107
0.388	1	0.999632488055862	0.981800808526277	0.988533627342889
0.398	1	0.999726443768997	0.985577507598784	0.989392097264438
0.408	1	0.999875637358537	0.988434274343987	0.992196244248228
0.418	1	0.999872822078087	0.989364746280046	0.993037008775277

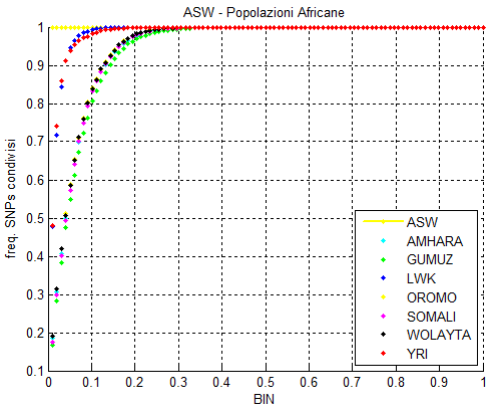
BIN	ASW	AMHARA	CEU	TSI
0.429	1	0.999901532805987	0.98948042143959	0.993271408409098
0.439	1	0.999933080153246	0.991149850266843	0.994696602144781
0.449	1	0.999846211681077	0.993575065787225	0.99603567888999
0.459	1	0.999876360039565	0.993800339126749	0.99597286986011
0.469	1	1	0.995022584359224	0.997325303338943
0.48	1	1	0.995303441696632	0.997633374917443
0.49	1	0.999981234048942	0.996359405494671	0.998254766551569
0.5	1	0.999981093549119	0.996710277546699	0.997428722680178
0.51	1	1	0.997534081796311	0.998396150761828
0.52	1	1	0.997632916863924	0.998538583455118
0.531	1	1	0.998381162619573	0.998864711447493
0.541	1	1	0.998401693341109	0.99894166180695
0.551	1	1	0.998783534267405	0.999152818507657
0.561	1	1	0.999097371093309	0.999458422655986
0.571	1	1	0.998932279243054	0.999386628501329
0.582	1	1	0.999049495769097	0.999513156369538
0.592	1	1	0.998700961288646	0.999881905571695
0.602	1	1	0.999473822678243	0.999928248547033
0.612	1	1	0.999658827829901	0.999829413914951
0.622	1	1	0.999799579116144	0.999924842168554
0.633	1	1	0.99980102967145	0.999925386126794
0.643	1	1	0.999765055994988	0.999947790221108
0.653	1	1	0.999920140552627	1
0.663	1	1	0.99989311386046	1
0.673	1	1	0.9999730043463	1
0.684	1	1	0.999972601238424	0.999972601238424
0.694	1	1	0.999917378132746	0.999972459377582
0.704	1	1	0.999915855608224	0.999971951869408
0.714	1	1	0.999972399326544	1
0.724	1	1	0.999972191323693	1
0.735	1	1	0.999971760187512	0.999971760187512
0.745	1	1	1	1
0.755	1	1	0.999971504288605	1
0.765	1	1	0.999941593902403	0.999970796951202
0.776	1	1	1	1
0.786	1	1	1	1
0.796	1	1	1	1
0.806	1	1	1	1
0.816	1	1	1	1
0.827	1	1	1	1
0.837	1	1	1	1
0.847	1	1	1	1
0.857	1	1	1	1
0.867	1	1	1	1
0.878	1	1	1	1
0.888	1	1	1	1

BIN	ASW	AMHARA	CEU	TSI
0.898	1	1	1	1
0.908	1	1	1	1
0.918	1	1	1	1
0.929	1	1	1	1
0.939	1	1	1	1
0.949	1	1	1	1
0.959	1	1	1	1
0.969	1	1	1	1
0.98	1	1	1	1
0.99	1	1	1	1
1	1	1	1	1

Tabella 3.1: Dati di quattro popolazioni in relazione agli Afroamericani

Si noti che le frequenze in tutte le colonne si saturano, cioè ad un certo punto, diverso per ogni popolazione, compaiono solo degli 1: le mutazioni ad alta frequenza nella popolazione di riferimento sono interamente condivise dalle altre popolazioni.

I precedenti dati si apprezzano meglio se rappresentati sul piano cartesiano, ponendo *BIN* sull'asse delle ascisse e le frequenze delle varie popolazioni sull'asse delle ordinate.
I seguenti quattro grafici sono stati realizzati con i dati dell'esempio precedente, completi delle 19 popolazioni, suddivise per continente di appartenenza (non si dispone di dati dell'Oceania). In ogni grafico viene riportata poi la popolazione di riferimento, ASW, che corrisponde ad un segmento orizzontale in corrispondenza di 1.



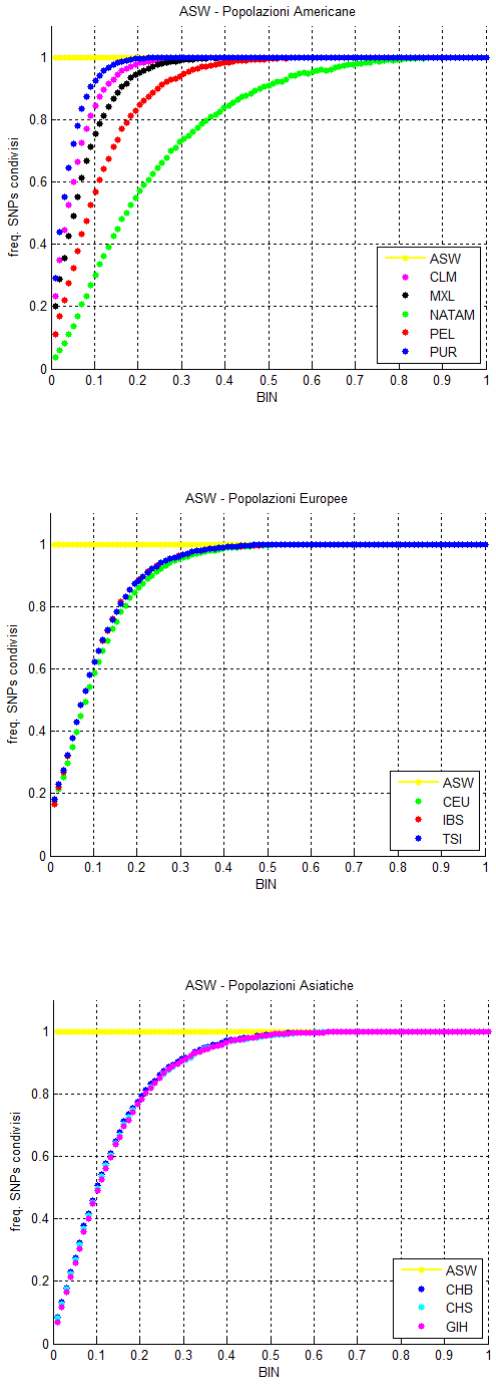


Figura 3.1: Rappresentazione grafica dei dati in relazione ad ASW

Utilizzando i dati aventi qualsiasi altra popolazione come riferimento si ottengono dei grafici del tutto simili a quelli di Figura 3.1.

Si potrebbe condurre un'analisi dei dati sulle 19 matrici contemporaneamente, ma nessun risultato sarebbe apprezzabile; si è così deciso di utilizzarne solo alcune. Poiché il continente di appartenenza fornisce una comoda diversificazione tra le popolazioni, è stata scelta una popolazione di riferimento per ognuno dei quattro continenti. Esse sono:

- YRI per l'Africa,
- ASW per l'America,
- CHB per l'Asia,
- CEU per l'Europa.

In questo modo si dovranno analizzare solo quattro matrici di dati, che rappresentano i quattro continenti in relazione agli altri.
La ragione per la quale sono state scelte tali popolazione è data dal fatto che la loro storia demografica è ben conosciuta, dunque qualsiasi parametro che potesse essere determinato matematicamente, può essere confrontato con i valori attesi in base alle nostre conoscenze. In particolare di queste popolazioni si può affermare che:

- CEU e CHB si sono separati circa 60 mila anni fa da YRI;
- CEU e CHB si sono separati circa 40 mila anni fa;
- CEU e YRI si sono mescolati (in proporzioni rispettivamente del 20% e dell'80%) per creare gli ASW, circa 300 anni fa.

3.2 Ricerca della funzione migliore per l'interpolazione dei dati del *progetto*

L'analisi dei dati è partita dall'osservazione dei grafici creati, relativi alle quattro popolazioni e con la successiva formulazione di alcune ipotesi.

La prima ipotesi riguarda le curve che si formano, la cui legge potrebbe essere per tutte la stessa. Si è deciso di procedere con una loro interpolazione con diverse funzioni, utilizzando il coefficiente di determinazione R^2 per confrontarle e stabilire quale fosse la migliore.
Dapprima si è pensato che possibili funzioni interpolanti potessero essere le funzioni razionali, vale a dire funzioni scrivibili come rapporto fra due polinomi:

$$f(x) = \frac{P(x)}{Q(x)}.$$

Si è però da subito notato lo svantaggio di tali funzioni: per ottenere un buon R^2 in ogni interpolazione, cioè prossimo a 1, i gradi dei polinomi al numeratore e al denominatore risultavano diversi da curva a curva e questo avrebbe reso le funzioni difficilmente confrontabili fra loro.

Il modello migliore, con $R^2 \geq 0.9$ è risultato essere quello esponenziale della forma

$$y = ae^{bx} + c,$$

con entrambi i parametri a e b negativi e c molto vicino a 1.

Osservando i grafici delle quattro popolazioni di riferimento in relazione solo con se stesse, e in particolare considerando le popolazioni a coppie, è stata avanzata anche un'altra ipotesi relativa all'area sottostante la curva e la sua possibile correlazione con il tempo di separazione delle popolazioni. In particolare sembra che minore sia l'area sottostante, più lontano nel tempo è avvenuto il momento di separazione delle due popolazioni, cioè la saturazione a 1 si osserva più tardi; viceversa, in due popolazioni separatesi di recente l'area sottostante la curva é maggiore e anzi molto vicina a 1, dovuta a una veloce convergenza ad 1 delle frequenze di mutazioni condivise.

3.3 Cluster Analysis e analisi discriminante

Ciò che non può essere apprezzato dai grafici cartesiani è la "somiglianza" tra le varie popolazioni, cioè a quali e quanto sia simile la popolazione di riferimento alle altre e viceversa; lo strumento utile a questo scopo è la *Cluster Analysis*, con la quale si raggruppano, in un insieme di dati, gli elementi omogenei secondo criteri di similarità.
Si è pensato così di condurre il clustering gerarchico delle quattro popolazioni di riferimento, una per volta, rispetto alle altre popolazioni, suddivise per continente di appartenenza. In questo modo si può apprezzare la somiglianza della popolazione di riferimento a quelle appartenenti allo stesso continente osservando il dendrogramma di output, il quale, riportando le distanze tra le varie popolazioni, permette di individuare, se esistono, quelle popolazioni più simili, date le loro mutazioni genetiche, a popolazioni di continenti diversi da quello di appartenenza.
Più precisamente, è stata considerata, ad esempio, la popolazione afroamericana (ASW) e quindi la matrice delle sue mutazioni condivise dalle altre popolazioni, senza la colonna di quelle condivise con se stessa; è stata trasposta, cosicché su ogni riga ci fosse una popolazione, e riordinata, riportando una dopo l'altra le popolazione di Africa, rinominata gruppo 1, America,

gruppo 2, Asia, gruppo 3, ed infine Europa, gruppo 4, ottenendo la matrice (riportata in parte) 3.2:

	Popolazione	**Gruppo**	**0,01**	**0,02**	**0,031**
AFRICA	AMHARA	1	0,185142	0,307346	0,406978
	GUMUZ	1	0,167323	0,282586	0,383112
	LWK	1	0,477262	0,716504	0,844158
	OROMO	1	0,191178	0,315239	0,419657
	SOMALI	1	0,17524	0,297642	0,401515
	WOLAYTA	1	0,191321	0,314204	0,418454
	YRI	1	0,480338	0,740271	0,858475
AMERICA	CLM	2	0,234324	0,349702	0,444493
	MXL	2	0,200095	0,286585	0,356712
	NATAM	2	0,036008	0,060484	0,083457
	PEL	2	0,112472	0,170264	0,22116
	PUR	2	0,29033	0,437843	0,550217
ASIA	CHB	3	0,085452	0,132517	0,179186
	CHS	3	0,082533	0,127981	0,173935
	GIH	3	0,06999	0,117895	0,165626
EUROPA	CEU	4	0,174401	0,213653	0,252788
	IBS	4	0,165182	0,220296	0,270111
	TSI	4	0,183263	0,22965	0,274216

Tabella 3.2: Dati relativi ad ASW per la Cluster Anaysis

Successivamente si è deciso di procedere con un altro strumento dell'analisi multivariata, avente un obbiettivo speculare a quello della Cluster Analysis: l'*Analisi discriminante*. Dati due o più gruppi definiti di individui e una serie di variabili rilevate su questi soggetti, che si pensa possano influenzare l'appartenenza a un gruppo oppure all'altro, l'Analisi discriminante ha lo scopo di trovare una modalità di assegnazione di un nuovo individuo a uno dei gruppi sulla base dei valori rilevati sulle altre variabili, dette discriminanti.
Nel nostro caso, con tale analisi si è voluto verificare se ognuna delle popolazioni esaminate, sempre in relazione a quelle di riferimento, sarebbe stata assegnata al proprio continente di appartenenza oppure se qualcuna si trovasse in un gruppo diverso da quello reale.
L'analisi discriminante consiste nelle seguenti operazioni: dati m gruppi, l'i-esimo di numerosità n_i, si cerca la direzione u che discrimini i dati nel miglior modo possibile, vale a dire che nella proiezione nella direzione u le varie popolazioni appaiono il più possibile ben separate fra di loro.
È probabile che i risultati di questa analisi, con i gruppi che si andranno a formare, possano fornire un interessante spunto per la riflessione sull'espansione dell'Homo sapiens sapiens fuori dall'Africa, sui tempi in cui ha raggiunto i vari continenti e sugli incroci tra diverse specie.

3.4 La simulazione matematica dei dati

Dopo l'osservazione dei dati reali rilevati con il *progetto 1000 genomi*, si è passati alla generazione di dati mediante un software di simulazione. Fornendo alcuni comandi al simulatore (verranno specificati in seguito), esso restituisce come output una matrice le cui righe rappresentano ognuna un individuo delle popolazioni generate e le colonne una mutazione; gli elementi sono soltanto 0 e 1 e significano rispettivamente assenza o presenza di una certa mutazione in un individuo. Come specificato nel primo capitolo infatti, le variabili generate casualmente sono binomiali, dunque assumono soltanto valore 1, in caso di successo, e 0, altrimenti.
Da tale matrice si calcolano le frequenze degli SNPs di una popolazione condivise con dalle altre popolazioni, ottenendo in questo modo i dati nella stessa forma di quelli ottenuti dalle popolazioni del *progetto 1000 genomi* già descritti.
In questa fase si è simulata sempre una popolazione iniziale con determinate caratteristiche, alla quale, in un certo momento, si impone una separazione in due diverse popolazioni. Trattandosi di simulazione coalescente però, è necessario interpretare il processo a ritroso, cioè: date due popolazioni iniziali esistenti nel presente con precise caratteristiche, si impone loro di unirsi, in un certo momento, in un'unica popolazione. Questo non è altro che il modello semplificato di ciò che è realmente accaduto nella storia per Europei e Asiatici, come già accennato, separatisi circa 40 mila anni fa da un'unica popolazione di provenienza africana. Come output si avranno le mutazioni genetiche delle due popolazioni di partenza.

3.4.1 Parte prima: variazione del tempo di unione

I comandi da fornire al simulatore per la generazione di popolazioni sono stati inizialmente decisi in modo da generare delle popolazioni simili a quelle di cui si dispongono i dati reali del progetto, e fatti variare uno per volta per comprendere in che modo emergessero nei grafici tali variazioni.
Le prime simulazioni sono state del tipo appena descritto, con tempo di separazione come parametro che varia. La riga di comando fornita al simulatore è la seguente:

```
-msms -N 10000 -ms 200 1 -t 5000 -r 4400 10000000 -I 2 100 100
                          0 -ej 0.1 2 1,
```

in cui:

- `-N` è la popolazione effettiva (generalmente indicata con N_e);

- `-ms` indica il numero degli aplotipi simulati (200) e il numero di repliche (1);
- `-t` è il tasso di mutazione (moltiplicato per $4N_e$ per il numero di paia base simulate);
- `-r` è il tasso di ricombinazione (moltiplicato per $4N_e$ per il numero di paia basi simulate) con la lunghezza degli aplotipi simulati (10000000);
- `-I` dà le indicazioni per le popolazioni che si vogliono simulare (2 popolazioni di 100 aplotipi l'una);
- `-ej` è il tempo in cui le popolazioni si uniscono (al tempo $0.1/4 * N_e$ la popolazione 2 si unisce alla popolazione 1).

Al parametro `-ej` sono stati assegnati 10 valori: $0.1, 0.2, 0.3, ..., 0.9, 1$ e per ognuno di questi sono state avviate 30 simulazioni; gli output delle simulazioni sono stati interpolati con la curva esponenziale $y = ae^{bx} + c$, rappresentati nello stesso piano cartesiano, in modo da osservare come variano le curve al variare del tempo `-ej`, e le costanti a, b, c salvate in un file di testo (.txt) per ognuno dei tempi, così da poter essere confrontate con alcuni test e analisi. Delle costanti a e b per ogni `-ej` sono state calcolate la media aritmetica e la deviazione standard. Tali medie sono state rappresentate sul piano cartesiano in funzione dei tempi di unione.

L'osservazione dei grafici delle medie ottenuti ha suggerito di procedere con alcuni test sui gruppi di dati relativi a ogni `-ej`: l'Analisi della varianza, il test di Bonferroni e l'analisi discriminante

Con l'Analisi della varianza si confrontano diversi gruppi di dati mediante il confronto della variabilità *interna* ai gruppi e *tra* gruppi; successivamente con il test di Bonferroni ne verranno approfonditi i risultati, poiché si calcoleranno le differenze tra le medie dei gruppi al fine di individuare quelle significative. Infine l'analisi discriminante permette di capire se i parametri a, b, c di un gruppo sono ben classificati o sarebbero più appropriati in altri gruppi.

Inoltre, sono state condotte delle trasformazioni e, specialmente sulle medie del parametro b, è stato osservato il loro andamento, dopo averlo rappresentato su un piano cartesiano, e calcolato il coefficiente di correlazione, affinché si potesse individuare l'esistenza di una dipendenza di b, o degli altri parametri, dal tempo di unione `-ej`.

3.4.2 Parte seconda: inserimento di un collo di bottiglia

Come già accennato, le simulazioni sono cominciate con il cercare di riprodurre quanto realmente accaduto alle popolazioni oggetto di studio del *progetto 1000 genomi*. Ad esempio, chiedendo al simulatore di generare due popolazioni che circa 300 anni fa si sono unite per crearne una terza, si riproduce l'origine della popolazione afroamericana, la quale ha avuto origine dall'incrocio di (parte della) popolazione europea e (di parte) di quella africana, a seguito della tratta degli schiavi africani.
Ciò che già era emerso in parte dall'osservazione dei grafici creati con i dati noti è l'asimmetria dei dati visibile dai grafici, cioè le mutazioni di una popolazione (popolazione 1) presenti in una seconda popolazione esaminata (popolazione 2) non sono le stesse della popolazione 2 presenti anche nella popolazione 1; in altre parole le curve interpolanti non sono uguali ma presentano a volte una notevole differenza nel valore delle aree sottostanti.
Questo porta a formulare alcune ipotesi su un probabile evento vissuto da solo una delle due popolazioni prima di unirsi all'altra come un collo di bottiglia. Si è provato così a riprodurre questa situazione aggiungendo alcune caratteristiche alle popolazioni generate, cambiando la riga di comando nel modo seguente:

```
msms -N 10000 -ms 200 1 -t 5000 -r 4400 10000000 -I 2 100 100 0
          -n 2 0.1 -g 2 10 -en 0.01 2 0.1 -ej 0.01 2 1
```

cioè aggiungendo:

- `-n`: indica, in questo caso, che la popolazione 2 è numerosa quanto 1/10 della popolazione 1 in origine (ciò è giustificato dal fatto che la popolazione 1 rappresenta l'Africa, il cui effettivo di popolazione è molto più grande di tutti gli altri continenti);
- `-g`: indica una continua crescita della popolazione 2 (le popolazioni non-africane si sono espanse molto di più di quanto non abbiano fatto quelle africane);
- `-en`: indica che al tempo 0.01, la popolazione 2 viene ridotta al 10%.

Dunque in questo modo la popolazione 2 (che nell'esempio precedente corrisponde alla popolazione europea) è meno numerosa della popolazione 1 ma sta crescendo, fino a quando, nello stesso momento dell'unione con la popolazione 1, subisce un collo di bottiglia che ne riduce la dimensione.
Il file di output viene in questo modo utilizzato due volte, una con la popolazione 1 come riferimento e un'altra con la popolazione 2. Di conseguenza i

dati sono stati rappresentati e interpolati con le solite curve esponenziali per capire se l'evento aggiunto è sufficiente a spiegare l'asimmetria osservata.

Capitolo 4

Risultati

4.1 L'interpolazione dei dati del *progetto*

L'interpolazione dei dati reali del *progetto 1000 genomi* è stata effettuata con l'utilizzo del software *Matlab* e ha fornito i risultati riportati nelle tabelle seguenti. Si tratta dei valori che assumono i parametri a, b, c nelle curve esponenziali con il relativo coefficiente di correlazione R^2, per ognuna delle popolazioni di riferimento in relazione alle altre.
Come in precedenza, le popolazioni sono state raggruppate per continente di appartenenza.
Si noti che è stata riportata anche l'interpolazione dei dati delle popolazioni di riferimento in relazione a se stesse, ma come si era già osservato, non si tratta di un'esponenziale bensì di una retta orizzontale in corrispondenza di 1, dunque a e b sono sempre nulli e $c = 1$.

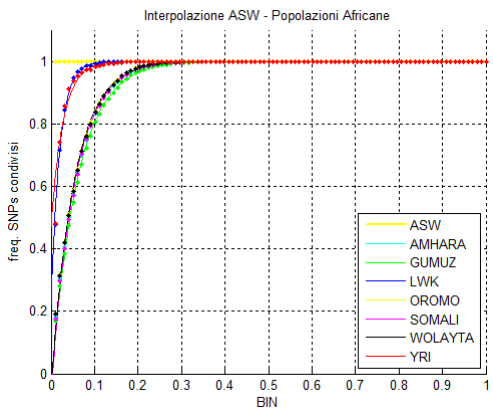
Popolazione di riferimento: ASW (Afroamericani)

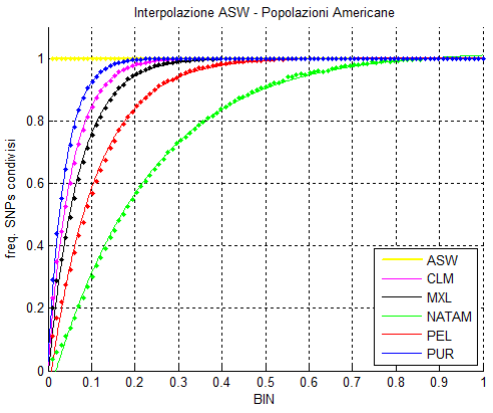
Africa	**a**	**b**	**c**	R^2
AMHARA	-1,0521	-18,3473	1,0002	0,9969
GUMUZ	-1,0414	-16,7222	1,0002	0,9982
LWK	-0,7731	-49,5153	1	0,9918
OROMO	-1,0493	-18,919	1,0001	0,9969
SOMALI	-1,0657	-18,4979	1,0001	0,9971
WOLAYTA	-1,0467	-18,7457	1,0001	0,997
YRI	-0,5276	-36,1312	1	0,9274

America	**a**	**b**	**c**	R^2
ASW	0	0	1	
CLM	-0,9948	-18,6769	1,0001	0,9967
MXL	-1,0204	-14,4257	1,0003	0,995
NATAM	-1,1095	-4,4564	1,023	0,9981
PEL	-1,0816	-9,5263	1,0016	0,9954
PUR	-0,9701	-25,3309	1	0,9976

Asia	**a**	**b**	**c**	R^2
CHB	-1,0962	-8,2306	1,0027	0,9962
CHS	-1,0948	-8,0216	1,003	0,9965
GIH	-1,1138	-8,0691	1,0032	0,996

Europa	**a**	**b**	**c**	R^2
CEU	-1,0533	-9,8332	1,0016	0,9917
IBS	-1,0519	-10,6979	1,001	0,9932
TSI	-1,0438	-10,6215	1,0012	0,9921





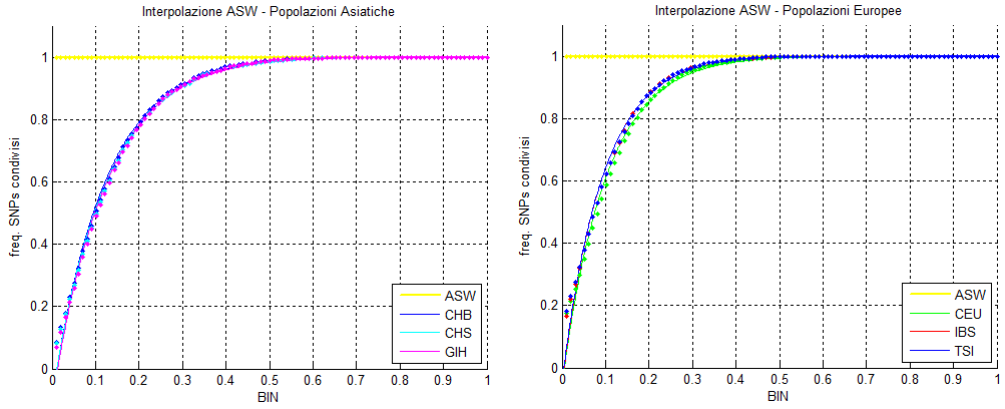


Figura 4.1: Funzioni interpolanti i dati in relazione ad ASW

Popolazione di riferimento: CEU (Europei)

Africa	a	b	c	R^2
AMHARA	-0,7697	-21,0565	0.9999	0.9935
GUMUZ	-0.6690	-8,0896	0.9967	0.9728
LWK	0,6713	-14,5408	0,9994	0,9824
OROMO	-0,7569	-20,4615	0,9998	0,9922
SOMALI	-0,7561	-16,3993	0,9997	0,9915
WOLAYTA	-0,7515	-18,5269	0,9998	0,9917
YRI	-0,6379	-10,4687	0,9979	0,9923

America	a	b	c	R^2
ASW	-0,7943	-32,572	1	0,9982
CLM	-0,8714	-62,766	1	0,9996
MXL	-0,8698	-54,5033	1	0,9991
NATAM	-0,9507	-5,1066	1,0094	0,9991
PEL	-0,871	-21,1309	0,9999	0,9977
PUR	-0,8388	-72,0234	1	0,9994

Asia	a	b	c	R^2
CHB	-0,8042	-12,461	0,9994	0,9935
CHS	-0,8011	-11,2213	0,9994	0,9935
GIH	-0,8777	-16,0099	0,9999	0,9962

Europa	**a**	**b**	**c**	R^2
CEU	0	0	1	
IBS	-0,9217	-73,4917	1	0,998
TSI	-1,0606	-118,347	1	0,9987

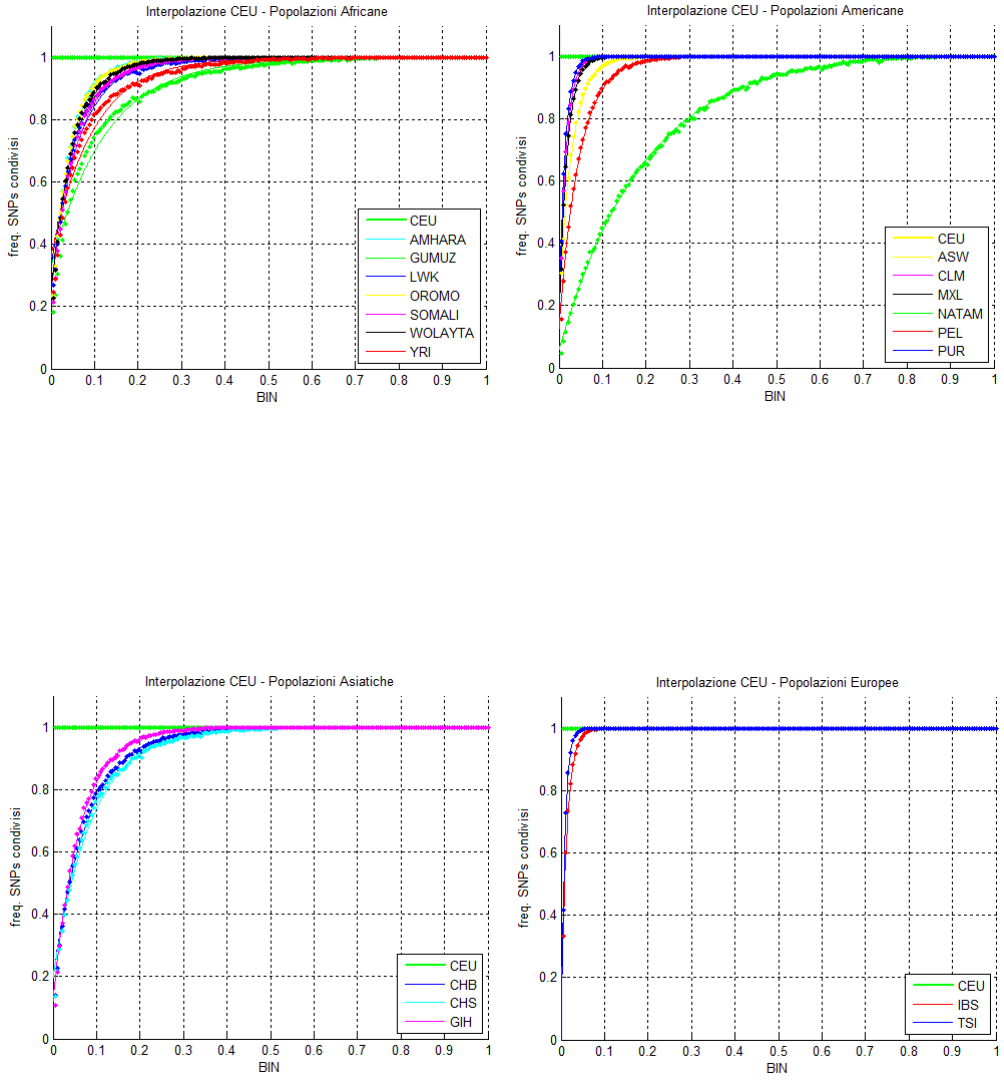


Figura 4.2: Funzioni interpolanti i dati in relazione a CEU

Popolazione di riferimento: CHB (Cinesi del Nord)

Africa	**a**	**b**	**c**	R^2
AMHARA	-0,6344	-11,0297	0,9980	0,9665
GUMUZ	-0,6074	-6,8057	0,9941	0,9566
LWK	-0,5844	-10,7595	0,9978	0,9569
OROMO	-0,6308	-10,9602	0,9982	0,9665
SOMALI	-0,6280	-9,6192	0,9976	0,9642
WOLAYTA	-0,6261	-10,6869	0,9979	0,9645
YRI	-0,5758	-8,8001	0,9958	0,9525

America	**a**	**b**	**c**	R^2
ASW	-0,6454	-15,8152	0,9994	0,9767
CLM	-0,6851	-19,1774	0,9997	0,9857
MXL	-0,7004	-20,338	0,9998	0,987
NATAM	-0,9085	-5,5095	1,0072	0,9979
PEL	-0,7331	-13,9986	0,9996	0,9874
PUR	-0,6759	-19,3511	0,9997	0,9854

Asia	**a**	**b**	**c**	R^2
CHB	0	0	1	
CHS	-1,3466	-126,1287	1	0,9981
GIH	-0,7337	-10,9818	0,9989	0,9829

Europa	**a**	**b**	**c**	R^2
CEU	-0,6513	-13,2286	0,9991	0,9768
IBS	-0,654	-12,1927	0,999	0,977
TSI	-0,6455	-13,8922	0,9991	0,9763

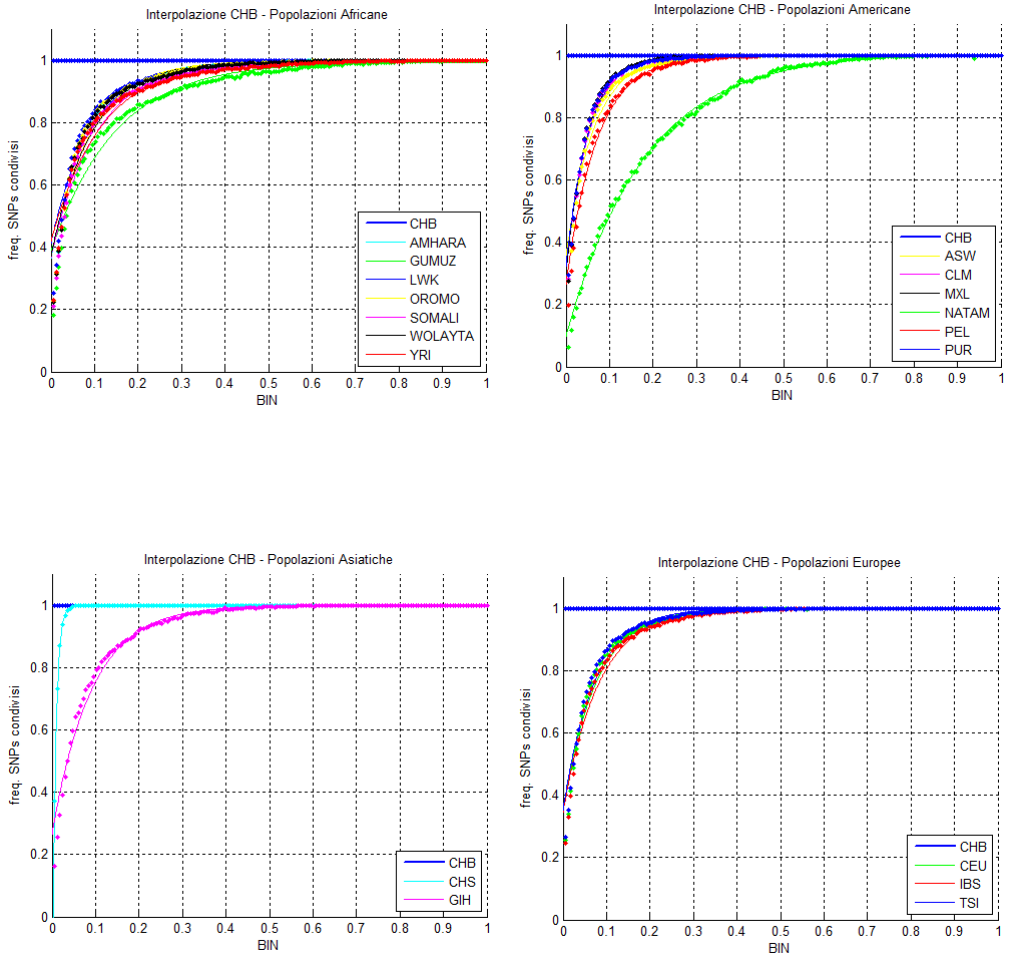


Figura 4.3: Funzioni interpolanti i dati in relazione a CHB

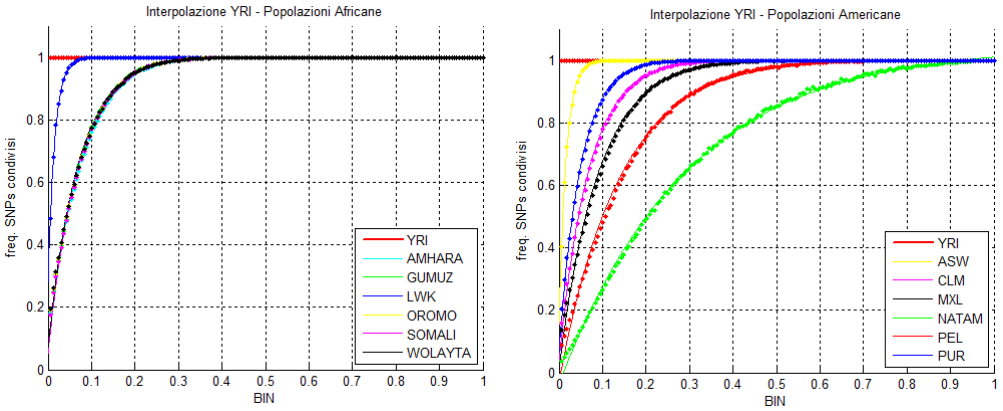
Popolazione di riferimento: YRI (Nigeriani)

Africa	a	b	c	R^2
AMHARA	-0,9458	-14,117	1,0004	0,9976
GUMUZ	-0,9459	-14,6285	1,0003	0,9982
LWK	-0,7186	-63,9387	1	0,9976
OROMO	-0,9347	-14,4744	1,0003	0,9979
SOMALI	-0,9484	-14,4353	1,0003	0,998
WOLAYTA	-0,9284	-14,3795	1,0003	0,9979
YRI	0	0	1	

America	a	b	c	R^2
ASW	-0,8428	-59,9383	1	0,9991
CLM	-0,9654	-14,7576	1,0002	0,9986
MXL	-0,9893	-11,1554	1,0007	0,9978
NATAM	-1,0821	-3,4073	1,0463	0,9988
PEL	-1,0326	-7,1879	1,004	0,9974
PUR	-0,9343	-20,4038	1,0001	0,999

Asia	a	b	c	R^2
CHB	-1,0352	-6,1606	1,0072	0,9973
CHS	-1,034	-6,0267	1,0077	0,9975
GIH	-1,0539	-5,9967	1,0084	0,9972

Europa	a	b	c	R^2
CEU	-1,0254	-7,3983	1,004	0,9967
IBS	-1,0171	-8,0807	1,0025	0,997
TSI	-1,0133	-8,0266	1,0029	0,9969



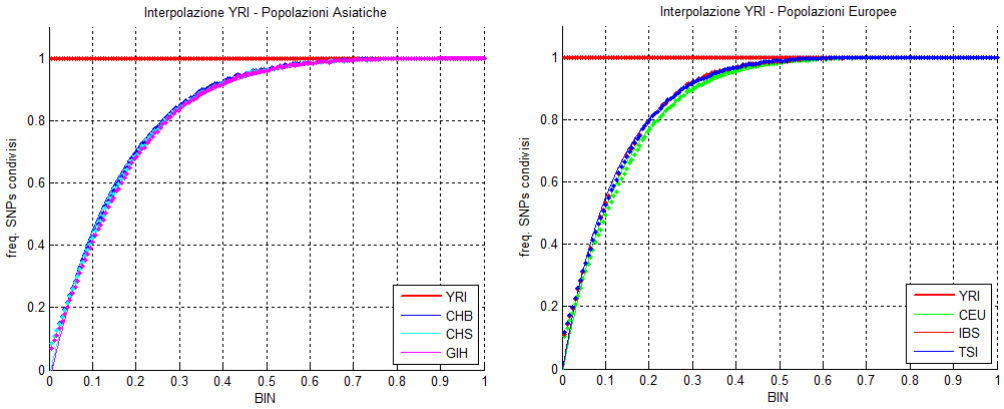


Figura 4.4: Funzioni interpolanti i dati in relazione ad YRI

4.2 Cluster Analysis e analisi discriminante

Entrambe le analisi, Cluster e discriminante, sono state effettuate con l'utilizzo del software *SPSS*. Ne vengono riportati gli output di seguito.

Cluster Analysis

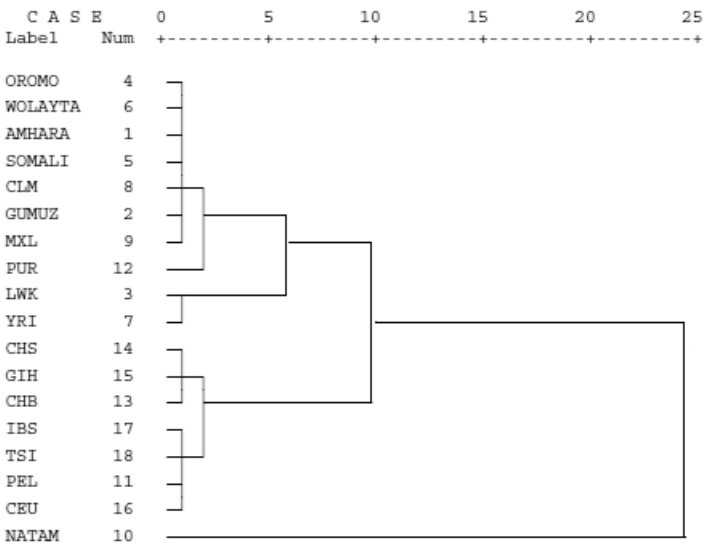


Figura 4.5: Cluster Analysis rispetto alla popolazione ASW

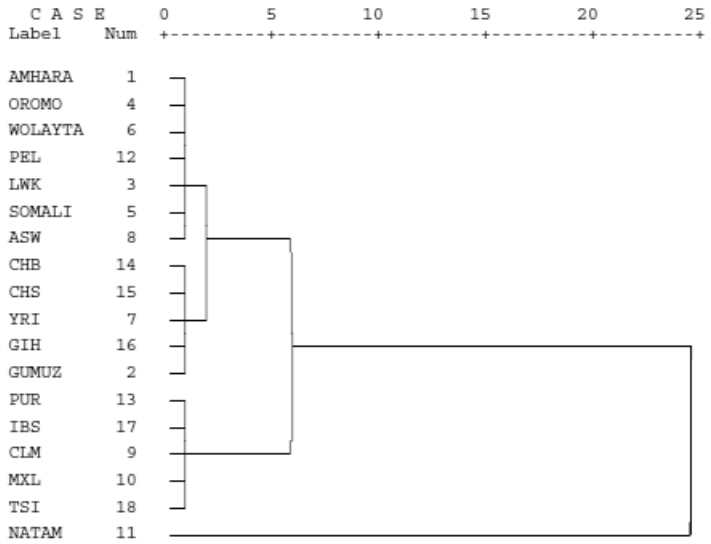


Figura 4.6: Cluster Analysis rispetto alla popolazione CEU

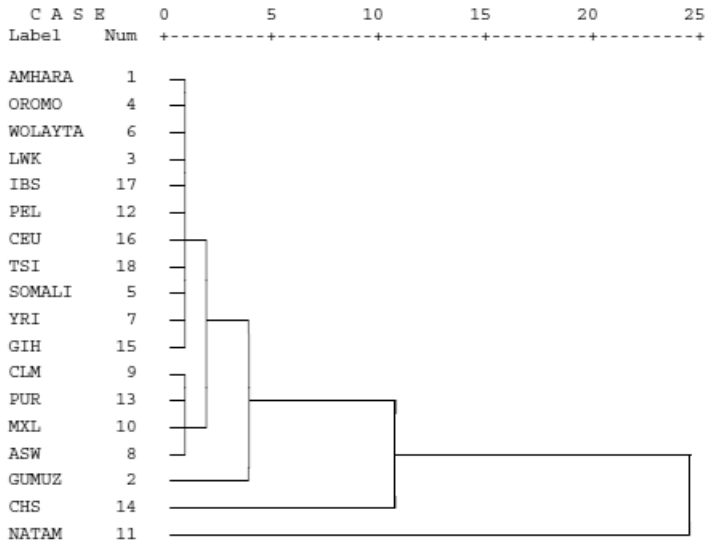


Figura 4.7: Cluster Analysis rispetto alla popolazione CHB

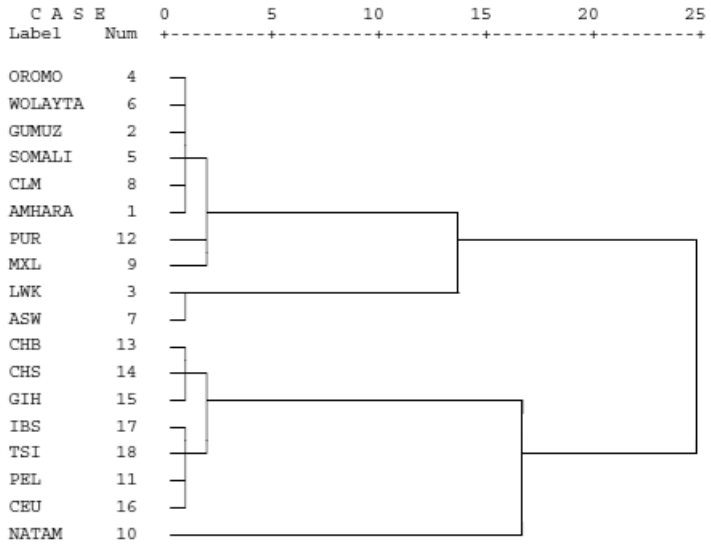


Figura 4.8: Cluster Analysis rispetto alla popolazione YRI

Analisi discriminante

Per quanto riguarda l'analisi discriminante, per ogni popolazione di riferimento vengono riportate:

- le tre funzioni discriminanti utilizzate, con i relativi autovalori e la percentuale di varianza spiegata;
- una tabella con i valori delle tre funzioni ai baricentri del gruppo;
- il grafico che rappresenta il piano trovato;
- la tabella che riassume i risultati della riclassificazione, nella quale cioè si evidenziano quante e quali popolazioni risultano ben classificate e quante invece sarebbero meglio in un altro continente.

ASW				
Funzione	Autovalore	% di varianza	% cumulata	Correlaz. canonica
1	15,007	96,5	96,5	0,968
2	0,536	3,4	100,0	0,591
3	0,005	0,0	100,0	0,067
CEU				
Funzione	Autovalore	% di varianza	% cumulata	Correlaz. canonica
1	6,517	66,7	66,7	0,931
2	2,981	30,5	97,2	0,865
3	0,274	2,8	100,0	0,464
CHB				
Funzione	Autovalore	% di varianza	% cumulata	Correlaz. canonica
1	7,464	61,4	61,4	0,939
2	4,187	34,5	95,9	0,898
3	0,501	4,1	100,0	0,578
YRI				
Funzione	Autovalore	% di varianza	% cumulata	Correlaz. canonica
1	1,253	59,2	59,2	0,746
2	0,855	40,4	99,7	0,679
3	0,007	0,3	100,0	0,085

Tabella 4.1: Autovalori

ASW			CEU		
Funzione			Funzione		
1	2	3	1	2	3
3,548	0,350	0,027	-1,692	-0,976	0,356
-0,841	-0,039	-0,095	1,626	-0,877	-0,494
-0,624	-1,358	0,044	-2,277	2,817	-0,344
-6,252	0,607	0,052	4,459	1,821	0,751
CHB			YRI		
Funzione			Funzione		
1	2	3	1	2	3
-2,915	-0,590	0,013	1,178	0,418	0,042
2,111	-0,456	-0,674	-0,996	0,722	-0,033
0,175	5,068	0,204	-0,895	-1,166	0,110
2,463	-1,089	1,182	0,531	-1,113	-0,127

Tabella 4.2: Funzioni ai baricentri di gruppo

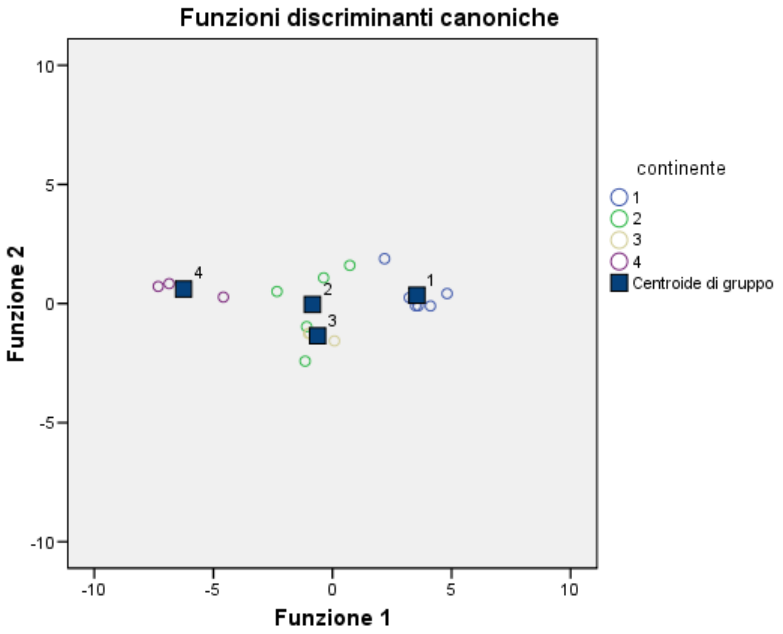


Figura 4.9: Funzioni discriminanti canoniche ASW

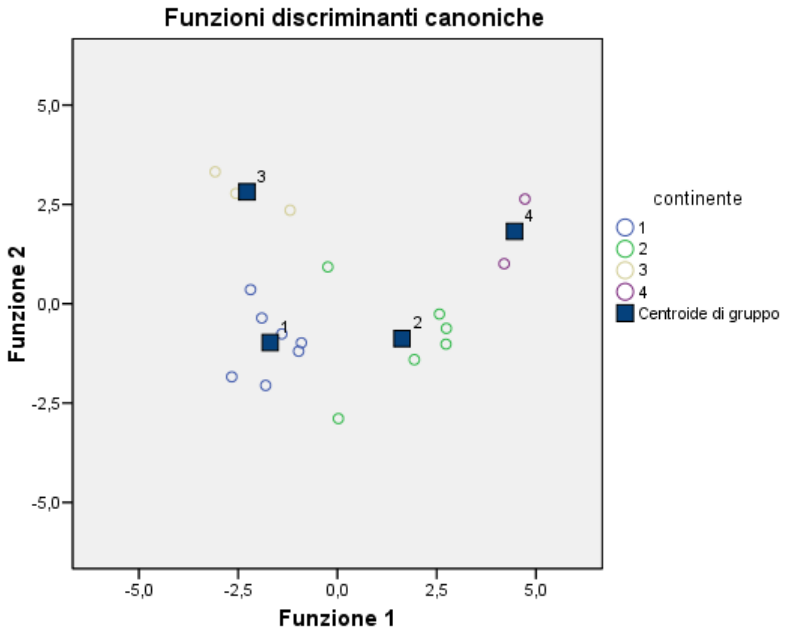


Figura 4.10: Funzioni discriminanti canoniche CEU

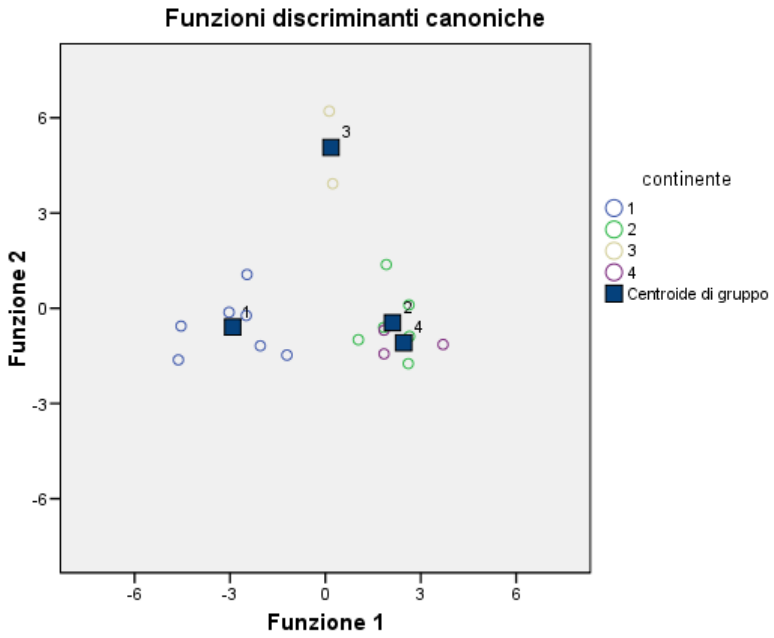


Figura 4.11: Funzioni discriminanti canoniche CHB

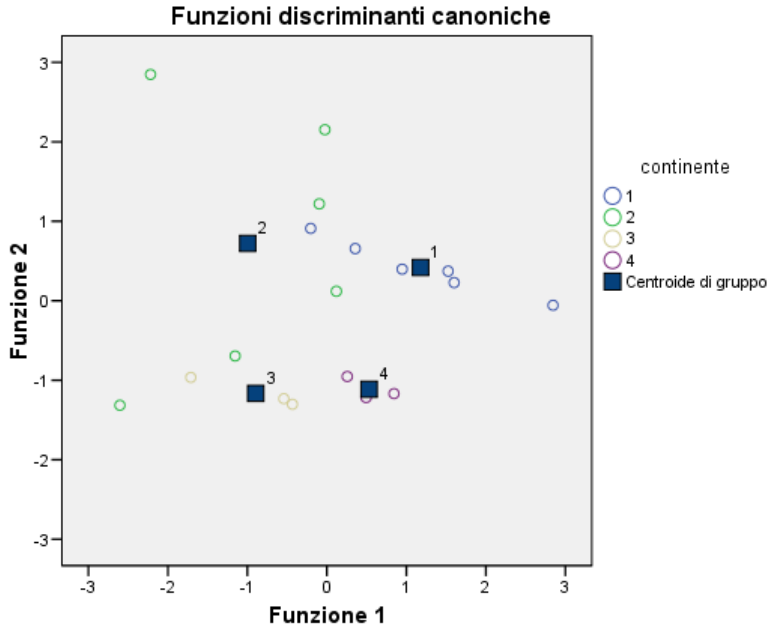


Figura 4.12: Funzioni discriminanti canoniche YRI

ASW						
	Continente	Gruppo di appartenenza previsto				Totali
		1	2	3	4	
Conteggio	1	7	0	0	0	7
	2	0	3	2	0	5
	3	0	0	3	0	3
	4	0	0	0	3	3
%	1	100	0,0	0,0	0,0	100
	2	0,0	60,0	40,0	0,0	100
	3	0,0	0,0	100	0,0	100
	4	0,0	0,0	0,0	100	100

Tabella 4.3: Risultati della classificazione rispetto ad ASW

CEU						
	Continente	Gruppo di appartenenza previsto				Totali
		1	2	3	4	
Conteggio	1	7	0	0	0	7
	2	0	6	0	0	6
	3	0	0	3	0	3
	4	0	0	0	2	2
%	1	100	0,0	0,0	0,0	100
	2	0,0	100	0,0	0,0	100
	3	0,0	0,0	100	0,0	100
	4	0,0	0,0	0,0	100	100

Tabella 4.4: Risultati della classificazione rispetto ad CEU

CHB						
	Continente	Gruppo di appartenenza previsto				Totali
		1	2	3	4	
Conteggio	1	7	0	0	0	7
	2	0	6	0	0	6
	3	0	0	2	0	2
	4	0	0	0	3	3
%	1	100	0,0	0,0	0,0	100
	2	0,0	100	0,0	0,0	100
	3	0,0	0,0	100	0,0	100
	4	0,0	0,0	0,0	100	100

Tabella 4.5: Risultati della classificazione rispetto ad CHB

YRI						
	Continente	Gruppo di appartenenza previsto				Totali
		1	2	3	4	
Conteggio	1	5	1	0	0	6
	2	1	3	2	0	6
	3	0	0	3	0	3
	4	0	0	0	3	3
%	1	83,3	16,7	0,0	0,0	100
	2	16,7	50,0	33,3	0,0	100
	3	0,0	0,0	100	0,0	100
	4	0,0	0,0	0,0	100	100

Tabella 4.6: Risultati della classificazione rispetto a YRI

Sono state omesse, per ragioni di spazio, le matrici di struttura per tutte le popolazioni, le quali indicano le variabili con cui le funzioni discriminanti sono correlate; da queste è scaturito che le variabili di ASW sono correlate con la funzione 2, quelle di CEU quasi esclusivamente con la funzione 3, quelle di CHB si dividono tra le tre funzioni e quelle di YRI quasi esclusivamente con la funzione 2.

4.3 La simulazione di popolazioni con tempo di unione `-ej` variabile

4.3.1 Rappresentazione delle simulazioni

I dati delle simulazioni, effettuate facendo variare il tempo di unione delle due popolazioni, sono stati interpolati ottenendo le seguenti curve esponenziali, rappresentate per ognuno dei 10 tempi `-ej` utilizzati, prima nello stesso piano poi separatamente, in modo da poterne apprezzare l'andamento al variare del parametro.

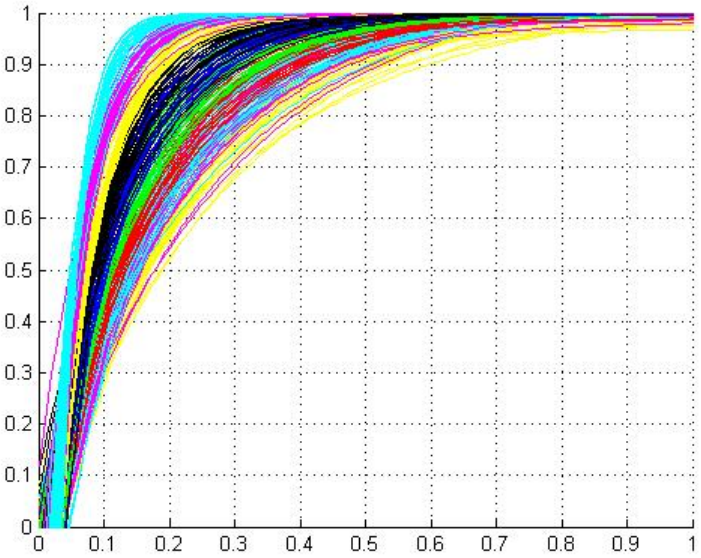


Figura 4.13: Rappresentazione delle interpolazioni delle simulazioni

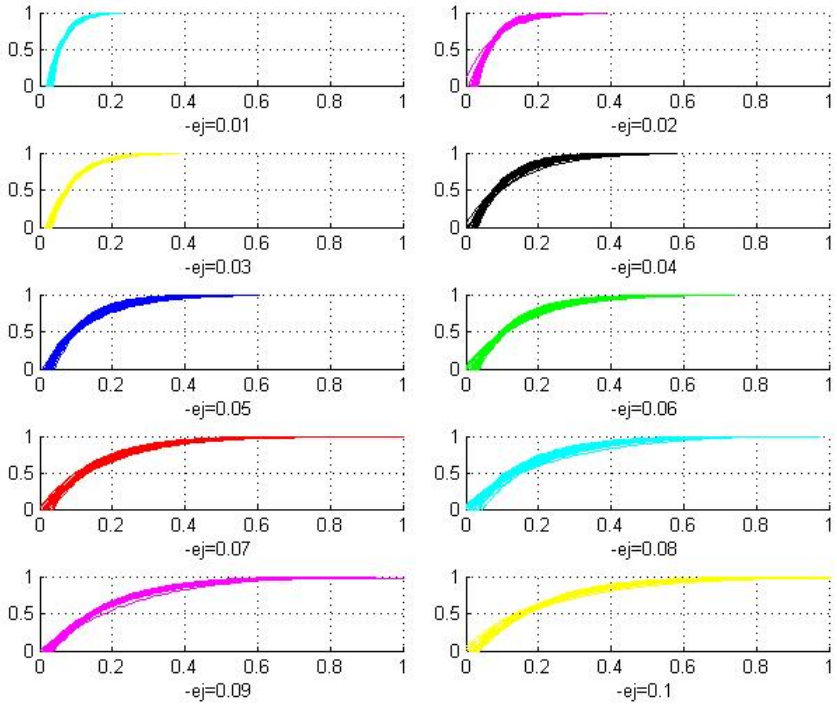


Figura 4.14: Rappresentazione delle interpolazioni per ogni tempo

4.3.2 I parametri a, b e c

Per ogni tempo -ej dei parametri a, b e c la media aritmetica e la deviazione standard che sono state calcolate vengono riportate nella tabella seguente:

-ej	Media a	St.Dev.a	Media b	St.Dev.b	Media c	St.Dev.c
0,01	-2,2556	0,4223	-27,3819	2,9189	1,0001	0,0000
0,02	-1,6532	0,3428	-18,8732	2,9602	1,0001	0,0002
0,03	-1,4637	0,1473	-14,564	1,1953	1,0003	0,0004
0,04	-1,2675	0,1655	-11,096	1,7272	1,0013	0,0012
0,05	-1,2473	0,1366	-9,6131	1,2527	1,0014	0,0018
0,06	-1,1798	0,1186	-8,0171	1,0785	1,0031	0,0034
0,07	-1,1306	0,0939	-6,7673	0,7663	1,0047	0,0057
0,08	-1,1372	0,115	-6,0692	0,953	1,0112	0,0095
0,09	-1,1419	0,0745	-5,4991	0,6106	1,0121	0,0114
0,1	-1,1266	0,0644	-4,9677	0,4572	1,0126	0,0121

Tabella 4.7: Medie e deviazioni standard dei parametri a, b, c

Successivamente a e b sono stati rappresentati sul piano cartesiano in funzione del tempo (c è stato tralasciato poiché approssimabile con 1 in ogni simulazione dunque indipendente dal tempo -ej).

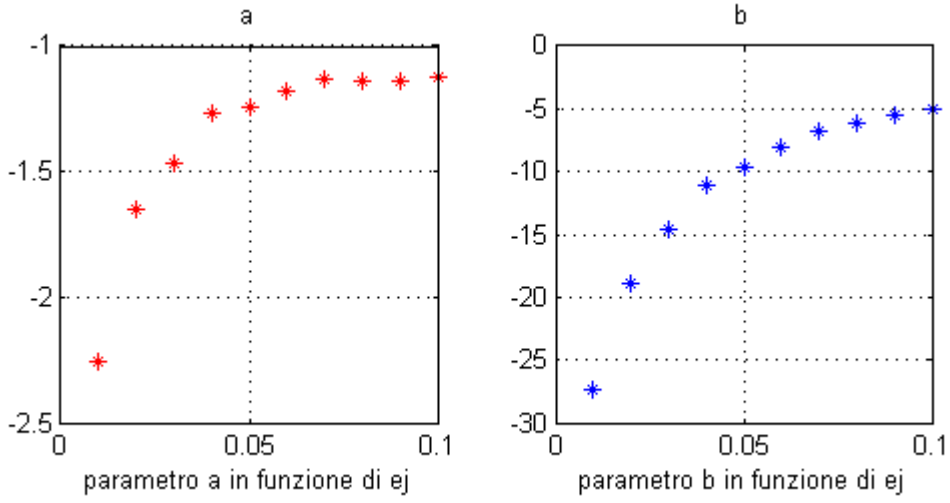


Figura 4.15: Rappresentazione delle medie in funzione di -ej

I dati a disposizione sono stati utilizzati per condurre l'analisi della varianza, test di Bonferroni e analisi discriminante; vengono riportati i risultati

nelle tabelle successive.
Si noti che i valori assunti da `-ej` (che sono $0.01, 0.02, ..., 0.1$) vengono indicati nelle tabelle con $1, 2, ..., 10$; questo è stato necessario per svolgere tutte le analisi citate, nel momento in cui si inserivano i dati nel software *SPSS*.

Analisi della varianza

	Varianza	Somma dei quadrati	df	Media dei quadrati	F	Sig.
a	Fra gruppi	34,803	9,000	3,867	95,138	0,000
	Entro gruppi	11,909	293,000	0,041		
	Totale	46,713	302,000			
b	Fra gruppi	13911,546	9,000	1545,727	586,992	0,000
	Entro gruppi	771,557	293,000	2,633		
	Totale	14683,104	302,000			
c	Fra gruppi	0,008	9,000	0,001	19,958	0,000
	Entro gruppi	0,012	293,000	0,000		
	Totale	0,020	302,000			

Tabella 4.8: ANOVA

Test di Bonferroni

Variabile dipendente	(I)	(J)	Differenza fra medie (I-J)	Sig.
a	1	2	-,6023533*	,000
		3	-,7918433*	,000
		4	-,9880267*	,000
		5	-1,0083052*	,000
		6	-1,0757471*	,000
		7	-1,1249733*	,000
		8	-1,1183500*	,000
		9	-1,1136733*	,000
		10	-1,1289858*	,000
	2	1	,6023533*	,000
		3	-,1894900*	,014
		4	-,3856733*	,000
		5	-,4059518*	,000
		6	-,4733938*	,000
		7	-,5226200*	,000
		8	-,5159967*	,000
		9	-,5113200*	,000
		10	-,5266325*	,000
	3	1	,7918433*	,000
		2	,1894900*	,014
		4	-,1961833*	,009
		5	-,2164618*	,002
		6	-,2839038*	,000
		7	-,3331300*	,000
		8	-,3265067*	,000
		9	-,3218300*	,000
		10	-,3371425*	,000
	4	1	,9880267*	,000
		2	,3856733*	,000
		3	,1961833*	,009
		5	-0,02	1,000
		6	-0,09	1,000
		7	-0,14	,404
		8	-0,13	,578
		9	-0,13	,738
		10	-0,14	,302
	5	1	1,0083052*	,000
		2	,4059518*	,000
		3	,2164618*	,002
		4	0,02	1,000
		6	-0,07	1,000
		7	-0,12	1,000
		8	-0,11	1,000
		9	-0,11	1,000
		10	-0,12	,859

Variab. dipend.	(I)	(J)	Differenza fra medie (I-J)	Sig.
a	6	1	1,0757471*	,000
		2	,4733938*	,000
		3	,2839038*	,000
		4	0,09	1,000
		5	0,07	1,000
		7	-0,05	1,000
		8	-0,04	1,000
		9	-0,04	1,000
		10	-0,05	1,000
	7	1	1,1249733*	,000
		2	,5226200*	,000
		3	,3331300*	,000
		4	0,14	,404
		5	0,12	1,000
		6	0,05	1,000
		8	0,01	1,000
		9	0,01	1,000
		10	0,00	1,000
	8	1	1,1183500*	,000
		2	,5159967*	,000
		3	,3265067*	,000
		4	0,13	,578
		5	0,11	1,000
		6	0,04	1,000
		7	-0,01	1,000
		9	0,00	1,000
		10	-0,01	1,000
	9	1	1,1136733*	,000
		2	,5113200*	,000
		3	,3218300*	,000
		4	0,13	,738
		5	0,11	1,000
		6	0,04	1,000
		7	-0,01	1,000
		8	0,00	1,000
		10	-0,02	1,000
	10	1	1,1289858*	,000
		2	,5266325*	,000
		3	,3371425*	,000
		4	0,14	,302
		5	0,12	,859
		6	0,05	1,000
		7	0,00	1,000
		8	0,01	1,000
		9	0,02	1,000

(*)La differenza tra le medie è significativa al livello 0.05.

Tabella 4.9: Test di Bonferroni del parametro a

Variabile dipendente	(I)	(J)	Differenza fra medie (I-J)	Sig.
b	1	2	-8,50875133*	,000
		3	-12,82	,000
		4	-16,29	,000
		5	-17,77	,000
		6	-19,36	,000
		7	-20,61	,000
		8	-21,31	,000
		9	-21,88	,000
		10	-22,41	,000
	2	1	8,50875133*	,000
		3	-4,30922367*	,000
		4	-7,77715400*	,000
		5	-9,26006980*	,000
		6	-10,86	,000
		7	-12,11	,000
		8	-12,80	,000
		9	-13,37	,000
		10	-13,91	,000
	3	1	12,81797500*	,000
		2	4,30922367*	,000
		4	-3,46793033*	,000
		5	-4,95084613*	,000
		6	-6,54691613*	,000
		7	-7,79668433*	,000
		8	-8,49475567*	,000
		9	-9,06487000*	,000
		10	-9,59622613*	,000
	4	1	16,28590533*	,000
		2	7,77715400*	,000
		3	3,46793033*	,000
		5	-1,48291580*	,019
		6	-3,07898580*	,000
		7	-4,32875400*	,000
		8	-5,02682533*	,000
		9	-5,59693967*	,000
		10	-6,12829580*	,000
	5	1	17,76882113*	,000
		2	9,26006980*	,000
		3	4,95084613*	,000
		4	1,48291580*	,019
		6	-1,59607000*	,006
		7	-2,84583820*	,000
		8	-3,54390954*	,000
		9	-4,11402387*	,000
		10	-4,64538000*	,000

Variabile dipendente	(I)	(J)	Differenza fra medie (I-J)	Sig.
b	6	1	19,36489113*	,000
		2	10,85613980*	,000
		3	6,54691613*	,000
		4	3,07898580*	,000
		5	1,59607000*	,006
		7	-1,25	,129
		8	-1,94783954*	,000
		9	-2,51795387*	,000
		10	-3,04931000*	,000
	7	1	20,61465933*	,000
		2	12,10590800*	,000
		3	7,79668433*	,000
		4	4,32875400*	,000
		5	2,84583820*	,000
		6	1,25	,129
		8	-0,70	1,000
		9	-1,27	,121
		10	-1,79954180*	,001
	8	1	21,31273067*	,000
		2	12,80397933*	,000
		3	8,49475567*	,000
		4	5,02682533*	,000
		5	3,54390954*	,000
		6	1,94783954*	,000
		7	0,70	1,000
		9	-0,57	1,000
		10	-1,10	,382
	9	1	21,88284500*	,000
		2	13,37409367*	,000
		3	9,06487000*	,000
		4	5,59693967*	,000
		5	4,11402387*	,000
		6	2,51795387*	,000
		7	1,27	,121
		8	0,57	1,000
		10	-0,53	1,000
	10	1	22,41420113*	,000
		2	13,90544980*	,000
		3	9,59622613*	,000
		4	6,12829580*	,000
		5	4,64538000*	,000
		6	3,04931000*	,000
		7	1,79954180*	,001
		8	1,10	,382
		9	0,53	1,000

(*)La differenza tra le medie è significativa al livello 0.05.

Tabella 4.10: Test di Bonferroni del parametro b

Variabile dipendente	(I)	(J)	Differenza fra medie (I-J)	Sig.
c	1	2	0,00	1,000
		3	0,00	1,000
		4	0,00	1,000
		5	0,00	1,000
		6	0,00	1,000
		7	0,00	,273
		8	-,01113433*	,000
		9	-,01208667*	,000
		10	-,01258403*	,000
	2	1	0,00	1,000
		3	0,00	1,000
		4	0,00	1,000
		5	0,00	1,000
		6	0,00	1,000
		7	0,00	,300
		8	-,01108033*	,000
		9	-,01203267*	,000
		10	-,01253003*	,000
	3	1	0,00	1,000
		2	0,00	1,000
		4	0,00	1,000
		5	0,00	1,000
		6	0,00	1,000
		7	0,00	,408
		8	-,01090500*	,000
		9	-,01185733*	,000
		10	-,01235470*	,000
	4	1	0,00	1,000
		2	0,00	1,000
		3	0,00	1,000
		5	0,00	1,000
		6	0,00	1,000
		7	0,00	1,000
		8	-,00993233*	,000
		9	-,01088467*	,000
		10	-,01138203*	,000
	5	1	0,00	1,000
		2	0,00	1,000
		3	0,00	1,000
		4	0,00	1,000
		6	0,00	1,000
		7	0,00	1,000
		8	-,00974643*	,000
		9	-,01069876*	,000
		10	-,01119613*	,000

Variabile dipendente	(I)	(J)	Differenza fra medie (I-J)	Sig.
c	6	1	0,00	1,000
		2	0,00	1,000
		3	0,00	1,000
		4	0,00	1,000
		5	0,00	1,000
		7	0,00	1,000
		8	-,00808385*	,000
		9	-,00903618*	,000
		10	-,00953355*	,000
	7	1	0,00	,273
		2	0,00	,300
		3	0,00	,408
		4	0,00	1,000
		5	0,00	1,000
		6	0,00	1,000
		8	-,00652000*	,005
		9	-,00747233*	,000
		10	-,00796970*	,000
	8	1	,01113433*	,000
		2	,01108033*	,000
		3	,01090500*	,000
		4	,00993233*	,000
		5	,00974643*	,000
		6	,00808385*	,000
		7	,00652000*	,005
		9	0,00	1,000
		10	0,00	1,000
	9	1	,01208667*	,000
		2	,01203267*	,000
		3	,01185733*	,000
		4	,01088467*	,000
		5	,01069876*	,000
		6	,00903618*	,000
		7	,00747233*	,000
		8	0,00	1,000
		10	0,00	1,000
	10	1	,01258403*	,000
		2	,01253003*	,000
		3	,01235470*	,000
		4	,01138203*	,000
		5	,01119613*	,000
		6	,00953355*	,000
		7	,00796970*	,000
		8	0,00	1,000
		9	0,00	1,000

(*)La differenza tra le medie è significativa al livello 0.05.

Tabella 4.11: Test di Bonferroni del parametro c

Analisi discriminante

Funzione	Autovalore	% di varianza	% cumulata	Correlaz. canonica
1	50,723	99,0	99,0	,990
2	,456	,9	99,9	,560
3	,048	,1	100,0	,214

Tabella 4.12: Autovalori

	Funzione		
	1	2	3
c	,082	,745*	,662
a	,232	-,602	,764*
b	,595	-,398	,699*

Tabella 4.13: Matrice di struttura

età	Funzione		
	1	2	3
1	-15,753	,993	-,258
2	-8,642	-,374	,416
3	-3,877	-,547	,063
4	-,524	-,779	,137
5	1,564	-,628	-,247
6	3,282	-,492	-,224
7	4,650	-,340	-,199
8	5,494	,559	,209
9	6,382	,774	,080
10	7,029	,844	,037

Tabella 4.14: Funzioni ai baricentri

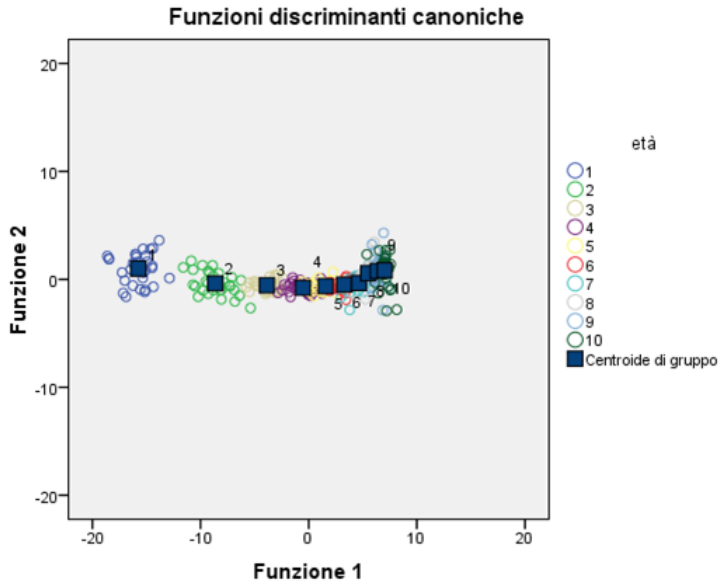


Figura 4.16: Funzioni discriminanti canoniche

	Età	Gruppo di appartenenza previsto										Tot.
		1	2	3	4	5	6	7	8	9	10	
Conteggio	1	30										30
	2		28	2								30
	3			30								30
	4			3	21	5	1					30
	5				6	20	5					31
	6					5	20	6				31
	7						6	19	5			30
	8						1	7	14	6	2	30
	9							3	4	15	8	30
	10							2	2	2	25	31
%	1	100	,0	,0	,0	,0	,0	,0	,0	,0	,0	100
	2	,0	93,3	6,7	,0	,0	,0	,0	,0	,0	,0	100
	3	,0	,0	100	,0	,0	,0	,0	,0	,0	,0	100
	4	,0	,0	10,0	70,0	16,7	3,3	,0	,0	,0	,0	100
	5	,0	,0	,0	19,4	64,5	16,1	,0	,0	,0	,0	100
	6	,0	,0	,0	,0	16,1	64,5	19,4	,0	,0	,0	100
	7	,0	,0	,0	,0	,0	20,0	63,3	16,7	,0	,0	100
	8	,0	,0	,0	,0	,0	3,3	23,3	46,7	20,0	6,7	100
	9	,0	,0	,0	,0	,0	,0	10,0	13,3	50,0	26,7	100
	10	,0	,0	,0	,0	,0	,0	6,5	6,5	6,5	80,6	100

Tabella 4.15: Risultati della classificazione

Capitolo 5

Discussione

5.1 L'interpolazione dei dati del progetto

Dall'osservazione dei grafici delle curve e dalla lettura dei relativi parametri a, b, c scaturiscono alcune considerazioni.
Innanzi tutto i coefficienti R^2 sono molto buoni, poiché tutti circa pari a 0.9, dunque le curve interpolanti si adattano quasi perfettamente ai dati.
Il parametro c, come era stato previsto, è circa 1 in ogni curva, essendoci la saturazione a 1 dei dati.
Il parametro b invece è tanto più grande in valore assoluto quanto prima avviene la saturazione a 1: quello più grande è $-126,1287$ di CHB in relazione a CHS, come ci si poteva facilmente aspettare; se invece la saturazione avviene più tardi, e dunque le mutazioni condivise con la popolazione di riferimento sono alte solo per alte frequenze, b assume valori più vicini a 0, come il $-3,4073$ di YRI in relazione a NATAM.
Infine dal parametro a dipende la radice della curva esponenziale: quanto più a è simile a 1, tanto più vicino all'origine è lo zero della funzione; se a è maggiore di -1, come il -0.5276 di ASW in relazione a YRI, allora l'intersezione della curva con l'asse delle x avviene a destra dell'origine, invece se a è minore di -1, come il $-1,3466$ di CHS in relazione a CHB, l'intersezione avviene a sinistra.

5.2 Cluster Analysis e Analisi discriminante dei dati del *progetto*

Cluster Analysis

In tutti dendrogrammi, prodotti con le Cluster Analysis condotte rispetto alle quattro popolazioni di riferimento, emerge innanzi tutto la grande distanza tra la popolazione NATAM da tutte le altre; essa rimane infatti esclusa da tutti le agglomerazioni di popolazioni, con l'eccezione del dendrogramma rispetto ad YRI, in cui si avvicina per somiglianza alle popopolazioni europee ed asiatiche.
Le agglomerazioni di popolazioni simili spesso rispettano i continenti di origine: ad esempio, il primo dendrogramma, rispetto ad ASW, mostra come la maggior parte delle popolazioni africane si siano raggruppate (solo YRI e LWK risultano leggermente più lontane), così come le popolazioni asiatiche e quelle europee; solo quelle americane si sono sparse tra le varie agglomerazioni (CLM, MXL, PUR tra le africane e PEL fra le europee). Nel secondo dendrogramma, rispetto a CEU, le popolazioni asiatiche sono ancora agglomerate tra loro con l'aggiunta però di due popolazioni africane (YRI e GUMUZ) che risultano più distanti dalle altre africane; lo stesso succede per le popolazioni europee, che risultano simili ad alcune popolazioni americane (PUR, CLM, MXL); dunque le popolazioni americane neanche in questo costituiscono un'agglomerazione a sè. Nel terzo grafico, rispetto a CHB, si nota una grande agglomerazione comprendente tutte le popolazioni africane (ad eccezione di GUMUZ), le tre europee, GIH e PEL; un'altra agglomerazione è costituita dalle quattro americane rimanenti, infine rimangono distanti da tutte le altre, insieme a NATAM, GUMUZ e CHS. Nell'ultimo grafico, rispetto ad YRI, le popolazioni asiatiche costituiscono una prima agglomerazione, le europee un'altra insieme a PEL; anche le africane si raggruppano, senza però LWK che risulta molto lontana dalle altre; infine le americane non sono agglomerate tra loro soltanto ma insieme alle africane; queste agglomerazioni possono essere facilmente distinte in due gruppi: il primo di africane e americane e il secondo tra europee e asiatiche.

Analisi discriminante

L'Analisi discriminante condotta successivamente alla Cluster Analysis ha permesso di vedere se si formano dei raggruppamenti di popolazioni corrispondenti ai continenti di appartenenza o se invece alcune popolazioni sarebbero da porre altrove.

Una prima idea la danno i grafici delle funzioni discriminanti canoniche e i valori ai baricentri di tali funzioni: si nota subito infatti come, rispetto ad ASW (Figura 4.9), mediante la funzione 1, le popolazioni europee siano discriminate dalle altre, mentre risultano tutte molto simili se si confrontano mediante la funzione 2. Rispetto a CEU (Figura 4.10), invece, le popolazioni non sono perfettamente raggruppate attorno al baricentro e si notano delle sovrapposizioni tra Africa e Asia, mediante alla funzione 1, tra Africa e America e tra Asia ed Europa mediante la funzione 2. Rispetto a CHB (Figura 4.11), mediante la funzione 1, le popolazioni americane ed europee sono leggermente sovrapposte; vale lo stesso per la funzione 2, mediante la quale neanche le popolazioni africane risultano discriminate; le popolazioni asiatiche invece, sono discriminate da tutte le altre mediante entrambe le funzioni. Infine, rispetto a YRI (Figura 4.12) le popolazioni risultano per nulla aggrlomerate attorno al baricentro e si sovrappongono mediante entrambe le funzioni discriminanti.

Dalla riclassificazione dei dati dell'analisi discriminante emerge che ci sia una sovrapposizione dei dati solamente se riferiti alle popolazioni ASW e YRI, mentre il 100% dei casi riferiti alle popolazioni CEU e CHB risulta ben classificato.
In particolare, in ASW solo il 60% delle popolazioni del continente americano risultano ben classificate poiché due di esse dovrebbero appartenere secondo l'Analisi all'Asia: si tratta di NATAM e PEL.
In YRI invece la sovrapposizione è ancora maggiore, infatti l'83% delle popolazioni africane risulta ben classificato, mentre una popolazione, GUMUZ, risulta appartenere all'America; inoltre, le popolazioni americane risultano ben classificate per il 50%, in quanto MXL risulta appartenere all'Africa mentre ancora NATAM e PEL risultano anche qui asiatiche.

5.3 La simulazione di popolazioni con tempo di unione -ej variabile

5.3.1 Rappresentazione delle simulazioni e osservazione dei parametri a e b

Dai grafici delle simulazioni si può constatare quanto già si era notato osservando le interpolazioni dei dati noti relativamente ai valori di b. Le curve infatti si saturano tanto prima quanto più b è grande in valore assoluto. Si può aggiungere che tale parametro è maggiore, sempre in valore assoluto, per

le simulazioni in cui il tempo di unione è piccolo, cioè l'unione è avvenuta (relativamente) poco tempo fa. All'aumentare di `-ej` la curva esponenziale si abbassa, indicando una saturazione più lenta.
Anche il valore del parametro a diminuisce in valore assoluto, avvicinandosi sempre di più a -1 all'allontanarsi del tempo di unione, così anche lo zero delle curve si avvicina all'origine da destra.
In generale si può anche notare dal grafico Figura 4.13 come siano più simili fra loro le curve interpolanti le simulazioni con `-ej` piccolo rispetto a quelle con `-ej` vicino a 1, indice di popolazioni aventi più mutazioni simili se separatesi poco tempo fa.

Dalla rappresentazione grafica delle medie di a e b emerge un andamento quasi esponenziale per entrambi in funzione del tempo `-ej`, con maggiore precisione per b. Questo risultato tornerà utile nel seguito della trattazione.

5.3.2 Analisi della varianza, test di Bonferroni e Analisi discriminante sui parametri a e b

Analisi della varianza

La tabella relativa all'Analisi della varianza (Figura 4.8) riporta, nella colonna *somma dei quadrati* i valori delle varianze fra gruppi SQQ_b, entro i gruppi SQQ_w e totale SQQ_{tot} per ognuno dei tre parametri; la colonna *df* indica i gradi di libertà relativi alle varianze mentre la *media dei quadrati* è il rapporto tra varianze e gradi di libertà; questi sono i valori necessari a calcolare F, il quale andrà confrontato con i valori della F di Fisher. Essendo il livello di significatività $\alpha = 0.05$ dalla colonna *Sig* si nota che l'Analisi è significativa per tutti tre i parametri, essendo i tre valori tutti inferiori alla soglia, e cioè si può concludere che le differenze tra le medie dei gruppi non hanno origine casuale.

Test di Bonferroni

Essendo risultata significativa l'Analisi per tutti tre i parametri, mediante il test di Bonferroni se ne approfondiscono i risultati.
Le Tabelle 4.9, 4.10 e 4.11 mostrano gli output di *SPSS* per il test di Bonferroni, condotte sui tre parametri. La prima colonna indica a quale parametro ci si riferisce, la seconda indica la variabile indipendente, cioè uno dei gruppi di simulazioni ad un determinato `-ej`, la terza il gruppo di simulazioni rispetto al quale viene calcolata la differenza, infine la differenza

tra le medie. Gli asterischi indicano i gruppi che rendono significativo il test F tra le varianze between e within, in corrispondenza dei quali *Sig* risulta inferiore a 0.05.
Per il parametro a, i primi tre gruppi sono significativamente diversi da tutti gli altri, e anche i gruppi dal quarto al decimo sono significativamente diversi dai primi tre, ma non da tutti gli altri. Per il parametro b, i gruppi dal terzo all'ultimo sono significativamente diversi da tutti gli altri (con solo qualche eccezione). Infine, per il parametro c sono gli ultimi tre gruppi ad essere significativamente diversi dagli altri.

Analisi discriminante

Mediante l'Analisi discriminante è stato possibile verificare se tutti tre i parametri delle simulazioni risultassero, congiuntamente, ben classificati. Come già fatto in precedenza, sono state riportate le tabelle con gli autovalori relativi alle tre funzioni canoniche (Tabella 4.12), la matrice di struttura (tabella 4.13), nella quale gli asterischi indicano con quale funzione discriminante ogni variabile è maggiormente correlata e in cui le variabili sono ordinate per dimensione assoluta crescente della correlazione entro la funzione. A seguire si trova la tabella dei valori delle tre funzioni canoniche nei baricentri dei tre gruppi (Tabella 4.14), rappresentati poi nella Figura 4.16. Osservando i valori dalla tabella delle funzioni ai baricentri, si può notare quanto sia diverso il valore della prima funzione in corrispondenza dei primi due gruppi e di conseguenza nel grafico quanto siano discriminati dagli altri gruppi. Diverso è invece il discorso per la seconda funzione, i cui valori sono simili tra loro, tutti attorno allo 0.
Dalla riclassificazione, infine, si ottiene il 73.3% dei casi correttamente classificati; in tutti i gruppi infatti, ad eccezione del primo e del terzo, si notano diversi casi di classificazione diversa da quella reale, arrivando ad un minino di solo il 46.7% di casi corretti nell'ottavo gruppo.
Questa Analisi, con la sua riclassificazione dei parametri, conferma quanto si era notato nei grafici iniziali delle simulazioni: all'allontanarsi del tempo di unione delle popolazioni, anche le curve di uno stesso gruppo di simulazioni risultavano distanziarsi tra loro sempre di più, tanto da rendere difficile capire a quale gruppo di simulazioni appartenessero (se non fosse per i colori).

5.4 Dalle simulazioni matematiche alla determinazione di un metodo per inferire il tempo di divisione di due popolazioni qualunque

5.4.1 Parte prima: calcolo del tempo della divisione di due popolazioni senza vincoli

Come già accennato, dall'osservazione dei grafici delle medie dei parametri a e b (Figura 4.15), ottenuti con la simulazione di due popolazioni generate dalla divisione di un'unica popolazione in certo numero di anni fa, si nota un loro andamento regolare. Sono state condotte alcune prove per capire se esiste una correlazione tra il tempo di divisione delle popolazioni `-ej` e le medie.
Per ogni media di a e b è stato calcolato il suo inverso, $\frac{1}{a}$ e $\frac{1}{b}$, e rappresentato in un grafico cartesiano, sempre in funzione di `-ej`, ottenendo il risultato seguente:

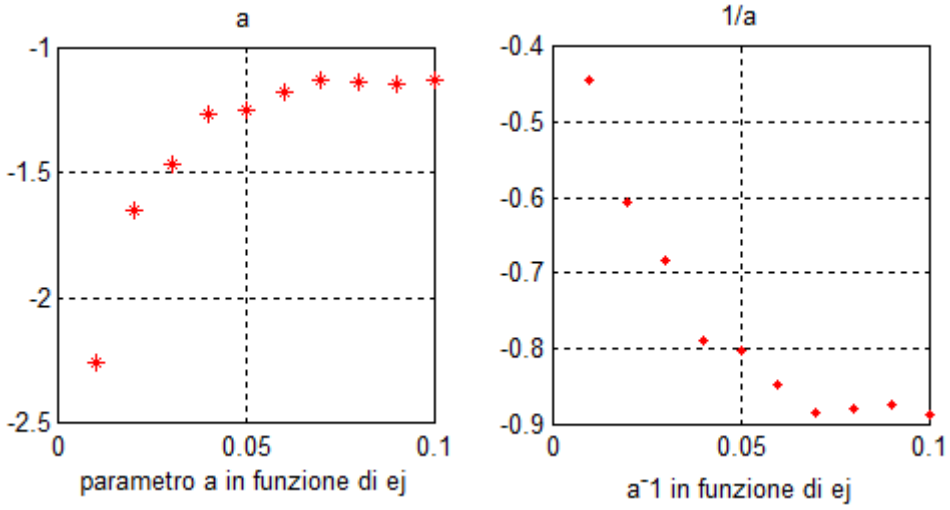


Figura 5.1: Parametro a e $\frac{1}{a}$ in funzione del tempo `-ej`

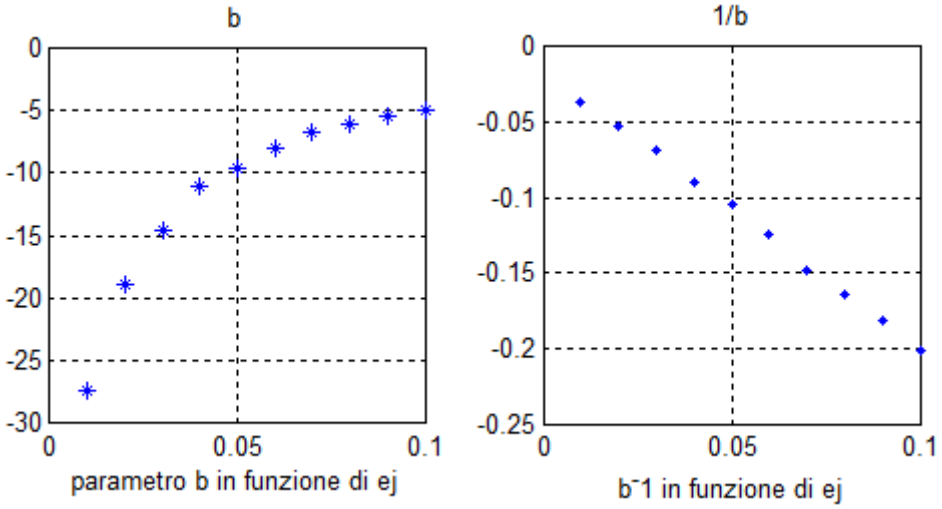


Figura 5.2: Parametro b e $\frac{1}{b}$ in funzione del tempo `-ej`

È evidente che vi sia una correlazione inversa tra il tempo e le medie e soprattutto come le medie inverse di b si dispongano lungo una retta decrescente. E' stato calcolato il coefficiente di correlazione lineare, indice che esprime proprio l'eventuale esistenza di relazione di linearità tra due variabili statistiche X e Y. Esso si indica con ρ_{XY} ed è definito come il rapporto tra la covarianza e le deviazioni standard delle due variabili:

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

e assume sempre valori compresi tra -1 e 1:

$$-1 \leq \rho_{XY} \leq 1.$$

La correlazione può essere di diversi tipi: diretta, inversa o nulla. Il segno di tale indice permette di comprendere di quale tipo di correlazione si è in presenza mentre il suo valore dà informazioni sull'entità della correlazione. Se:

- $\rho_{XY} > 0$, le due variabili sono direttamente correlate;
- $\rho_{XY} = 0$, le variabili non sono correlate;
- $\rho_{XY} < 0$, le variabili sono inversamente correlate.

Inoltre, per la correlazione diretta (e analogamente per quella inversa), si è stabilito convenzionalmente che se:

- $0 < \rho_{XY} \leq 0.3$, la correlazione è debole;
- $0.3 < \rho_{XY} \leq 0.7$, è moderata;
- $\rho_{XY} > 0.7$, è forte.

I coefficienti di correlazione ottenuti sono i seguenti:

$$\rho_{ej,a} = -0.8895, \quad \rho_{ej,b} = -0.9993,$$

dunque si può affermare con certezza che esista una correlazione in entrambi i casi inversa e forte, specialmente per b, con il tempo `-ej`.

Si è pensato di sfruttare questo fatto per inferire il tempo di divisione di una popolazione originale in due diverse popolazioni.
Per farlo si è trovata la migliore retta interpolante le medie inverse $\frac{1}{b}$, della forma:

$$\frac{1}{b} = Mej + Q,$$

cioè sono stati determinati i valori del coefficiente angolare M e dell'ordinata all'origine Q.
Si è esplicitata la variabile `-ej` di cui si vuole ricavare il valore. Infatti nella realtà, presi dei campioni di due popolazioni qualunque, di questi si conoscono le frequenze delle mutazioni di una condivise dall'altra, di conseguenza si possono calcolare le curve esponenziali interpolanti, cioè i parametri a, b (e c). Ciò che non è noto è proprio in tempo `-ej`.
I coefficienti M e Q trovati danno la retta:

$$\frac{1}{b} = -1.8541ej - 0.0153,$$

dalla quale

$$ej = \left(\frac{1}{b} + 0.0153\right)(-1.8541)^{-1}. \tag{5.1}$$

Oltre alle media dell'inverso di b sono state calcolate anche le relative deviazioni standard $\sigma_{\frac{1}{b}}$. Poiché questo indice esprime la dispersione dei dati attorno alla media aritmetica, è stata utilizzata per calcolare delle rette vicine a quella interpolante $1/b$ per ottenere un range di valori entro cui dovrebbe trovarsi l'effettivo `-ej` in un caso reale. Questo range di valori è stato ottenuto cercando la migliore interpolazione per i valori $\frac{1}{b} - \frac{\sigma_{\frac{1}{b}}}{2}$ e $\frac{1}{b} + \frac{\sigma_{\frac{1}{b}}}{2}$. Sono state ottenute rispettivamente le seguenti rette:

$$ej = \left(\frac{1}{b} + 0.0172\right)(-1.9608)^{-1}, \tag{5.2}$$

$$ej = \left(\frac{1}{b} + 0.0134\right)(-1.7473)^{-1}. \tag{5.3}$$

Per chiarezza la tabella seguente riassume i valori trovati a partire da b:

ej	b	1/b	std 1/b	std/2	1/b - std/2	1/b + std/2
0,01	-27,3819	-0,0365	0,0041	0,00205	-0,0386	-0,0345
0,02	-18,8732	-0,0530	0,0093	0,00465	-0,0576	-0,0483
0,03	-14,5640	-0,0687	0,0058	0,0029	-0,0716	-0,0658
0,04	-11,0960	-0,0901	0,0161	0,00805	-0,0982	-0,0821
0,05	-9,6131	-0,1040	0,0138	0,0069	-0,1109	-0,0971
0,06	-8,0171	-0,1247	0,0175	0,00875	-0,1335	-0,1160
0,07	-6,7673	-0,1478	0,0175	0,00875	-0,1565	-0,1390
0,08	-6,0692	-0,1648	0,0302	0,0151	-0,1799	-0,1497
0,09	-5,4991	-0,1818	0,0225	0,01125	-0,1931	-0,1706
0,1	-4,9677	-0,2013	0,0185	0,00925	-0,2106	-0,1921

Tabella 5.1: Valori ottenuti dal parametro b delle simulazioni

Nel seguente grafico sono state rappresentate le tre rette interpolanti

$$\frac{1}{b}, \qquad \frac{1}{b} - \frac{\sigma_{\frac{1}{b}}}{2}, \qquad \frac{1}{b} + \frac{\sigma_{\frac{1}{b}}}{2}$$

.

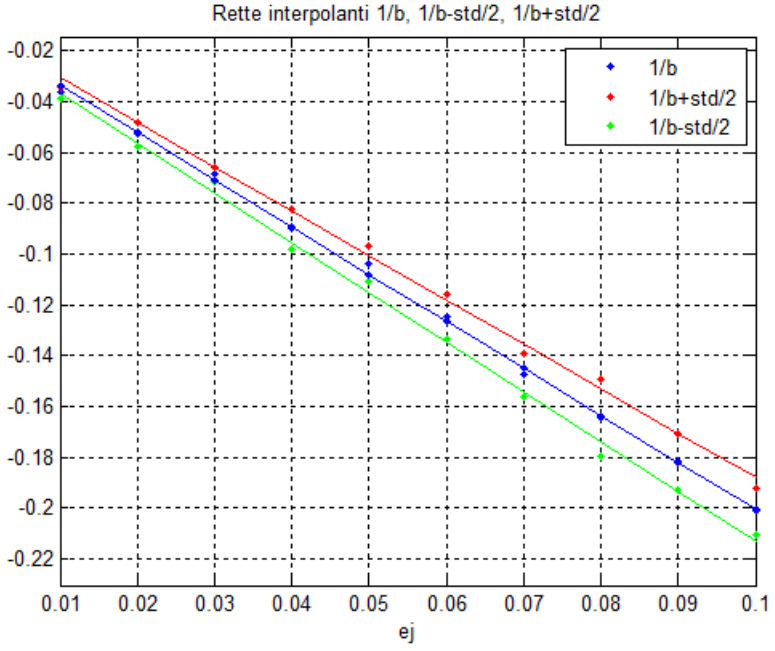


Figura 5.3: Rette interpolanti 1/b, 1/b-std/2, 1/b+std/2

Sono così state condotte diverse prove per constatare la validità di questo metodo. Cioè con il simulatore sono state generate due popolazioni che in un certo momento `-ej` si sono unite a formarne una unica. L'output è stato utilizzato come sempre per calcolare le mutazioni di una delle due popolazioni condivise dalla seconda, delle quali è stata individuata la migliore curva esponenziale interpolante. Il paramentro b di tale curva è stato utilizzato come variabile indipendente nelle tre rette interpolanti appena descritte e sono stati salvati tutti i dati che dovrebbero corrispondere ai tempi di unione delle popolazioni. Sono riportati nella tabella seguente.

ej reale	b	ej min	ej	ej max
0,013	-26,43252	0,0105	**0,0122**	**0,014**
	-26,85211	0,0102	**0,0118**	**0,0136**
0,025	-18,26208	0,0192	**0,0213**	**0,0237**
	-17,40873	0,0205	**0,0227**	**0,0252**
0,037	-12,8698	0,0309	0,0337	**0,0368**
	-12,47822	0,0321	**0,035**	**0,0382**
0,055	-9,03234	0,0477	**0,0515**	**0,0557**
	-8,67385	0,05	**0,0539**	**0,0583**
0,062	-7,807	0,0566	**0,0608**	**0,0656**
	-7,83736	0,0563	**0,0606**	**0,0654**
0,075	-6,93934	0,0647	0,0695	**0,0748**
	-6,3979	**0,0709**	**0,076**	0,0818
0,08	-5,94152	**0,0771**	**0,0825**	0,0887
	-6,50161	0,0697	0,0747	0,0804

Tabella 5.2: Prove di inferenza di alcuni `-ej`

Nella tabella sono stati evidenziati gli intervalli a cui appartiene il reale `-ej`. Si può facilmente notare che la precione del tempo inferito diminuisce con l'aumentare del tempo reale di divisione; è un risultato atteso poiché si tratta di tempi molto grandi e non è possibile individuare perfettamente il momento di divisione delle popolazioni. In generale però, le stime sono molto buone.

5.4.2 Altre osservazioni sulle simulazioni

Fino ad ora tutti i file output del simulatore sono stati trattati allo stesso modo, ossia sono state contate le frequenze delle mutazioni di una delle due popolazioni (1) condivise con l'altra popolazione (2), mai viceversa. Ci si

aspetta comunque che i risultati siano circa gli stessi contando le frequenze delle mutazioni della popolazione 2 condivise dalla popolazione 1, e dunque che le migliori curve esponenziali interpolanti i dati abbiano parametri a e b molto simili nei due casi. Questo dovrebbe accadere poiché nella riga di comando al simulatore, si richiede che la popolazione iniziale si suddivida a metà tra le due popolazioni, le quali continuano a esistere senza modifiche. Sono state condotte delle prove per avere conferma di quanto ci si aspetta; si riportano qui i grafici e le funzioni interpolanti di una di esse, realizzata con `ej` $= 0.062$.

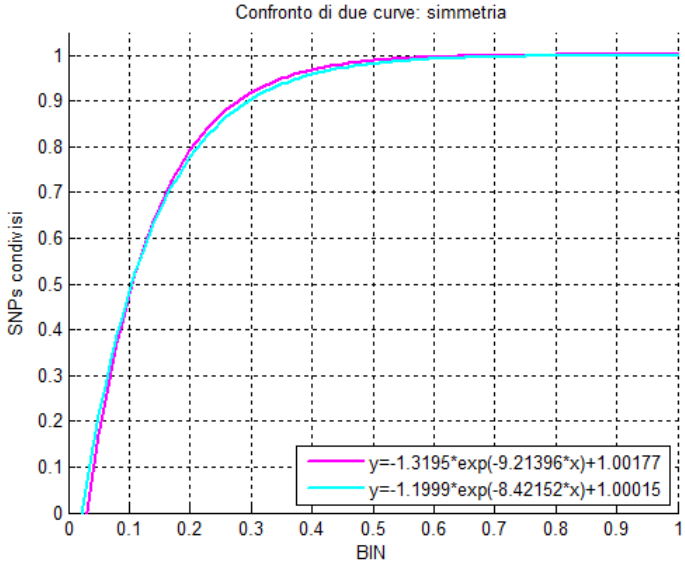


Figura 5.4: Confronto di due curve esponenziali: mutazioni della popolazione 1 condivise dalla popolazione 2 e viceversa

$$y_1 = -1.3195e^{-9.21396x} + 1.00177,$$

$$y_2 = -1.1999e^{-8.42152x} + 1.00015$$

Questi risultati sono però in contrasto con i dati reali: prese due popolazioni qualunque, le mutazioni di una condivise dall'altra non sono simmetriche, cioè non sono simili alle mutazioni della seconda popolazione condivise dalla prima, e di conseguenza non sono simili nemmeno le due curve esponenziali. Si veda, ad esempio, la curva dei dati di CEU in relazione ad ASW come differisce da quella di ASW in relazione a CEU:

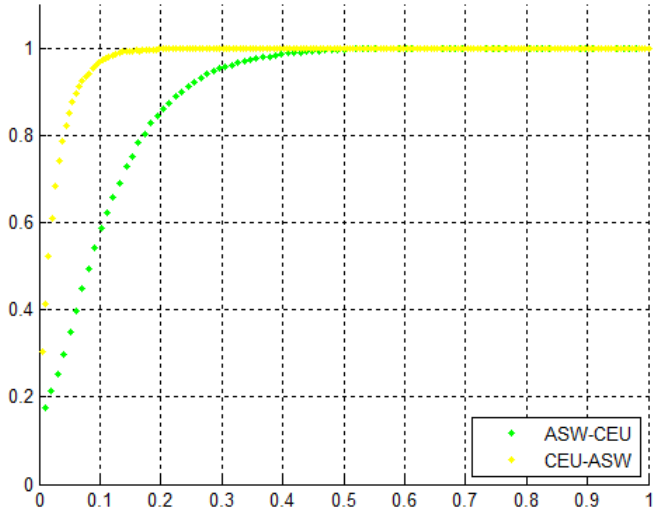


Figura 5.5: Esempio di asimettria dei dati

Tale asimmetria porta a pensare che una delle due popolazioni possa aver vissuto un particolare evento, come ad esempio un collo di bottiglia, subito dopo la separazione dall'altra popolazione. Sono così state simulate altre popolazioni con l'aggiunta di vincoli, analizzando l'output dai due "punti di vista".

5.4.3 Parte seconda: calcolo del tempo in presenza di un collo di bottiglia

Fissato il tempo di divisione nelle due popolazione, in queste simulazioni sono stati fatti variare altri due parametri: il terzo di `-en` e il secondo di `-n`. Più precisamente, fissato `-en` a 0.1, sono state effettuate 10 simulate per ognuno dei valori assunti da `-n`, da 0.1 a 1. Poi, fissato a `-en` a 0.2, si è ripetuta la stessa operazione, fino a `-en` pari a 1. In complesso le simulazioni generate sono 100.
Alla generazione di ogni simulazione, sono state calcolate le frequenze delle mutazioni prima rispetto a una popolazione poi rispetto all'altra; i dati trovati sono stati interpolati con delle esponenziali e di queste curve memorizzati i parametri a, b e c. Successivamente, di a e b sono state calcolate le medie aritmetiche, sempre rispetto ad entrambe le popolazioni, e salvate in dei file. Per ogni blocco di simulazioni con `-en` fissato si dispone di due file contenenti 10 medie ognuno.
Ad esempio, per `-en 0.01 2 0.2`, i due file delle medie contengono i pa-

rametri necessari a disegnare le dieci curve interpolanti di una popolazione rispetto all'altra e le 10 inverse. Questo è il grafico:

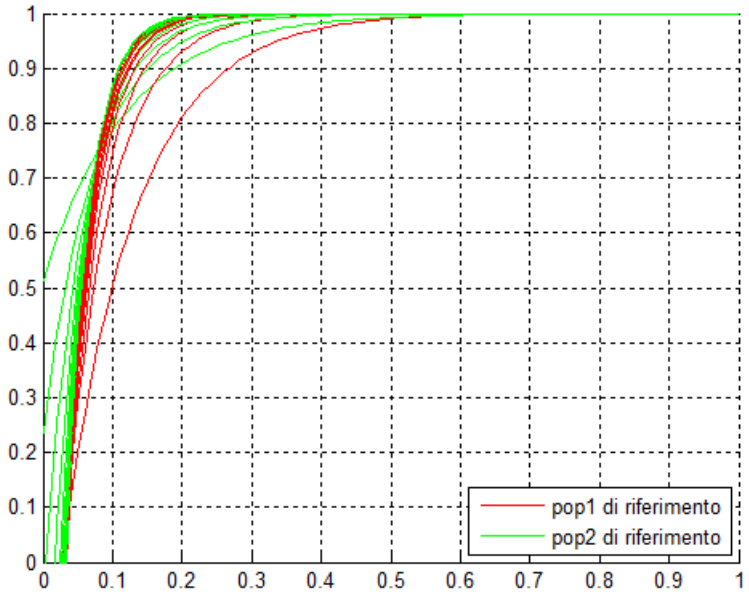


Figura 5.6: Grafico delle curve medie della popoalazione 1 rispetto alla 2 (rosse) e viceversa (verdi)

Uno strumento utile per confrontare tra loro le curve può essere la loro area sottostante. Si può infatti osservare l'andamento delle aree e trarre interessanti conclusioni. Si è così deciso di calcolarle per ognuno dei file contenenti i valori delle curve, come quelle rappresentati nella figura precedente, quindi di calcolare l'integrale:

$$\int_{x_0}^{1} ae^{bx} + 1\, dx, \tag{5.4}$$

in cui a e b sono i parametri contenuti nei file delle medie. È stato però necessario decidere l'estremo sinistro dell'intervallo su cui calcolarlo. Sarebbe stato naturale prendere lo 0, ma la scelta sarebbe risultata errata nei casi in cui le curve esponenziali intersecassero l'asse delle ascisse a destra dell'origine, poiché sarebbe stata considerata anche l'area compresa tra la curva e l'asse ma con segno negativo. Pertanto, per ogni curva è stato calcolato il valore della funzione all'origine: se questo risultava positivo, cioè la curva interseca l'asse x a sinistra dell'origine, si è posto $x_0 = 0$; se invece la funzione in 0 risultava negativa, è stato calcolato il punto di intersezione tra la curva e l'asse x, che si trovava necessariamente a destra, e utilizzato come estremo

sinistro nell'integrale.

Per i dati mostrati nella precedente figura sono state calcolate le seguenti aree, per ogni n le differenza tra le aree relative alle due curve e l'area media. Sono riportate nella tabella successiva e rappresentate nei grafici sottostanti.

- en 0,01 2 0,2				
n	aree rosse	aree verdi	differenza (v-r)	area media
0,1	0,8700	0,9419	0,0719	0,9060
0,2	0,9077	0,9434	0,0357	0,9256
0,3	0,9187	0,9397	0,0210	0,9292
0,4	0,9247	0,9373	0,0126	0,9310
0,5	0,9275	0,9367	0,0092	0,9321
0,6	0,9293	0,9356	0,0063	0,9325
0,7	0,931	0,9348	0,0038	0,9329
0,8	0,9323	0,9347	0,0024	0,9335
0,9	0,9329	0,9346	0,0017	0,9338
1	0,9342	0,9346	0,0004	0,9344

Tabella 5.3: Aree sottostanti le curve esponenziali, loro differenza e media.

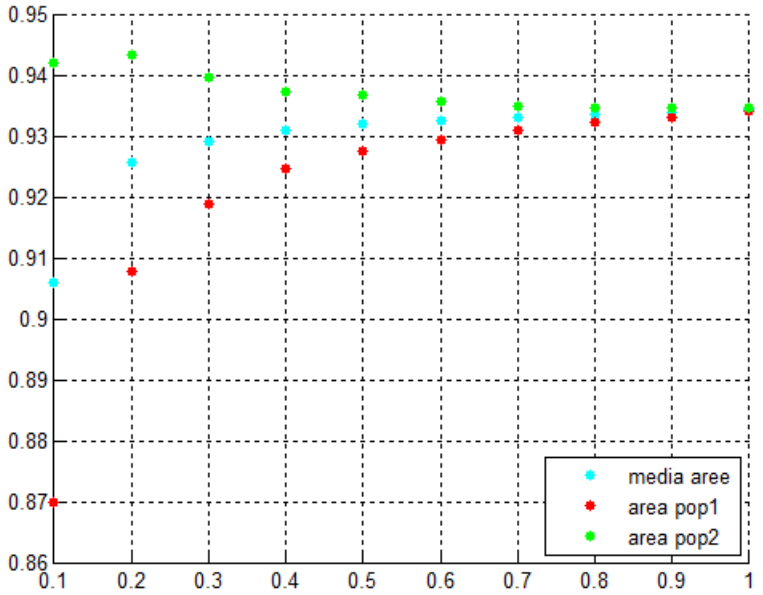


Figura 5.7: Aree sottostanti le curve e loro media

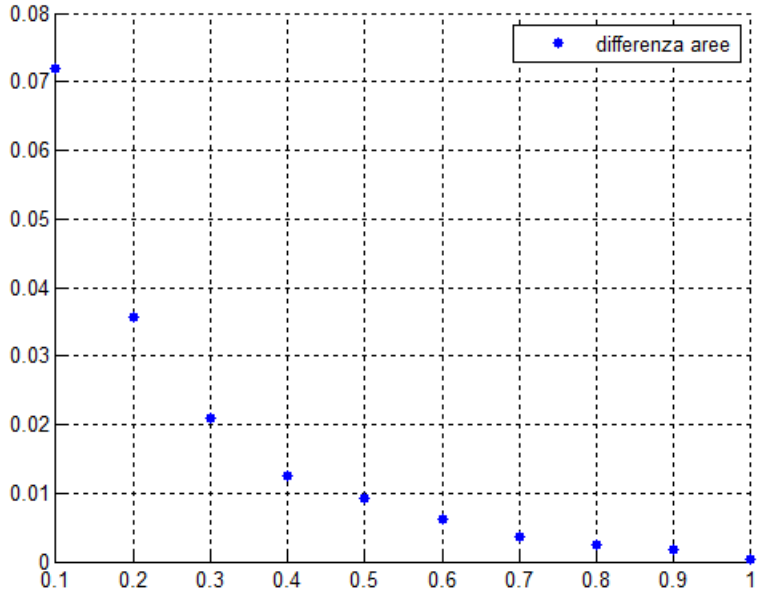


Figura 5.8: Differenza tra le aree sottostanti le curve

Si può facilmente notare che per `n` pari a 1 la differenza tra le medie è praticamente nulla, e questo vale per qualsiasi valore assegnato al terzo parametro di `-en`. Significa infatti che non vi è una popolazione più numerosa dell'altra quindi si è nelle stesse condizioni delle simulazioni della prima parte; così anche le curve sono simmetriche e non differiscono l'una dall'altra per l'area sottostante.
Inoltre, confrontando i grafici delle differenze di aree al variare del terzo parametro di `-en`, si è notato che tale parametro non è responsabile di notevoli differenze nelle curve, cioè le aree sottostanti e le loro medie sono risultate essere circa le stesse per tutti i valori di `-en`. Ad esempio, si veda come sono simili i valori di aree e loro differenze per `-en` pari a 0.01 e 0.06:

- en 0,01			**- en 0,06**			
aree rosse	aree verdi	v - r	aree rosse	aree verdi	v - r	**differenze**
0,8698	0,9429	0,0731	0,871	0,9439	0,0729	-0,0002
0,9067	0,9424	0,0357	0,9083	0,9425	0,0342	-0,0015
0,9189	0,9398	0,0209	0,9193	0,9401	0,0208	-0,0001
0,9242	0,9366	0,0124	0,9242	0,9373	0,0131	0,0007
0,9283	0,9374	0,0091	0,9279	0,9369	0,009	-0,0001
0,9295	0,9358	0,0063	0,9299	0,9361	0,0062	-0,0001
0,9312	0,935	0,0038	0,9309	0,9351	0,0042	0,0004
0,9323	0,9348	0,0025	0,9324	0,9352	0,0028	0,0003
0,9335	0,9345	0,001	0,9335	0,9345	0,001	0,0000
0,9341	0,9345	0,0004	0,9336	0,9343	0,0007	0,0003

Questo ha portato a pensare che potesse essere trovato un metodo per inferire il tempo di divisione delle due popolazioni senza che tale parametro influenzasse troppo il risultato.

Per poter calcolare il tempo di divisione, come nella parte precedente, è necessario conoscere il parametro b della curva esponenziale. Per il momento si sa che nei casi appena simulati, b assume due differenti valori, a seconda della popolazione che si tiene di riferimento per conteggiare le mutazioni condivise con l'altra. I due valori però coincidono quasi perfettamente quando `n` vale 1. Allora si è verificato che, per un qualsiasi `-en` fissato, l'area delle curve, quella rossa e quella verde dei grafici precedenti, ottenute con $n = 1$, sia equivalente alla media delle aree sottostanti le due curve esponenziali ottenute con un qualsiasi valore assegnato ad `n`. Cioè:

$$\forall\ n, \qquad \frac{(area\ rossa) + (area\ verde)}{2} \simeq\ Area_{n=1}.$$

Così questo valore medio può essere eguagliato all'integrale (5.4):

$$Area_{n=1} = \int_{x_0}^{1} ae^{bx} + 1\, dx. \tag{5.5}$$

Per ora nell'integrale i parametri a e b sono sconosciuti, ma sono quelli necessari per poter determinare il tempo `-ej`.
Si è deciso di sfruttare nuovamente le simulazioni della prima parte, delle quali erano stati memorizzati per ogni curva, al variare di `-ej`, a e b. Era stato notato che fossero linearmente correlati con `-ej`, ma non solo: essi sono anche direttamente correlati tra di loro. Il coefficiente di correlazione tra a e b risulta essere infatti 0.9824.
Si può dunque assumere che siano dipendenti l'uno dall'altro, pertanto si può trovare una retta che interpoli i dati e esprima al meglio la relazione esistente. Si è trovato che la migliore retta interpolante è:

$$a = 0.0492b - 0.8055.$$

Nel grafico seguente vengono riportati i loro valori, b sull'asse delle x e a sull'asse delle y.

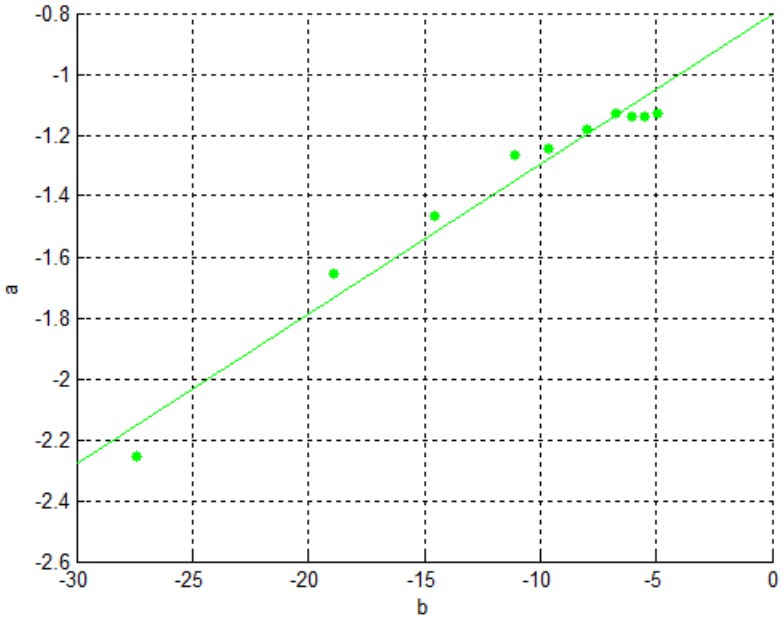


Figura 5.9: Parametri a e b, retta interpolante che ne eprime la relazione

L'espressione di a può essere inserita nell'integrale (5.5), in modo da rimanere con b unica incognita. Si ottiene infatti:

$$Area_{n=1} = \int_{x_0}^{1} (0.0492b - 0.8055)e^{bx} + 1\, dx$$

$$Area_{n=1} = \int_{x_0}^{1} 0.0492be^{bx} - 0.8055e^{bx} + 1\, dx$$

$$Area_{n=1} = \left[0.0492e^{bx} - \frac{0.8055}{b}e^{bx} + x\right]_{x_0}^{1}$$

$$Area_{n=1} - 1 + x_0 = (e^b - e^{bx_0})\left(0.0492 - \frac{0.8055}{b}\right). \qquad (5.6)$$

L'equazione (5.6) è quella che permette di ricavare il valore di b, necessario per risalire al tempo di divisione delle due popolazioni. Applicandolo a popolazioni reali, di queste si avranno a disposizione le mutazioni di una condivise dall'altra e viceversa e cioè le due curve esponenziali. Non si conosceranno altre informazioni, se una delle due ha vissuto un collo di bottiglia, ad esempio, o un altro evento. Non si sa nemmeno se ce ne sia una più numerosa dell'altra. L'avere a disposizione però le curve esponenziali interpolanti, e

quindi i loro parametri, con l'equazione (5.6) integrata ai risultati della prima parte di simulazioni, si riuscirà a risalire comunque ad una stima del tempo di divisione.
Di volta in volta, nell'utilizzarla, si è deciso di assegnare a x_0 il valore dell'ascissa del punto medio tra i punti di intersezione delle curve esponenziali con l'asse delle x. Per quei casi in cui una delle due curve intersecava l'asse delle x a sinistra dell'origine, nel calcolo della media delle ascisse è stato considerato il valore 0.
Ad $Area_{n=1}$ è stato invece assegnato il valore della media tra le aree sottostante le due esponenziali.

Per chiarezza si riassumono brevemente i passaggi di questo metodo. Siano date due popolazioni delle quali si vuole inferire il tempo di divisione; di esse si conoscono le equazioni (e quindi i parametri a e b) delle due curve esponenziali interpolanti i dati relativi alle mutazioni dell'una condivise dall'altra e viceversa. Tali informazioni vengono utilizzate per:

- calcolare le due aree sottostanti le curve esponenziali e la loro media, per sostituirla nell'equazione (5.6) ad $Area_{n=1}$;
- determinare i punti di intersezione delle due curve con l'asse delle x (se il punto è a sinistra dell'origine, quindi di ascissa negativa, si considera lo 0) e il loro punto medio, per sostituirlo ad x_0 nell'equazione (5.6);
- si determina il valore di b dall'equazione (5.6) (con l'utilizzo di un software);
- il valore trovato viene sostituito nell'equazione (5.1), (5.2) e (5.3) per trovare una stima dell'intervallo del tempo in cui dovrebbe essere avvenuta la divisione delle popolazioni.

Non ci si aspetta che tale metodo funzioni sempre correttamente, ma soltanto che sia utile a fornire una stima di `-ej`, anche nella realtà. Ci sono infatti notevoli limiti sui quali non è stato fino a qui possibile intervenire o impossibili da evitare. Primo fra tutti è il fatto che si stia lavorando con tempi molto lontani e grandi , dell'ordine di migliaia di anni.
Inoltre, parlando di questo secondo metodo appena presentato, è necessario aggiungere che si prevede che i risultati siano buoni solo per valori di `-n` maggiori di 0.03; per valori minori, infatti, si osserva un eccessivo distacco non regolare delle aree delle esponenziali dall'area della curva in corrispondenza di $\mathtt{n} = 1$, che rende imprecisi i risultati.

Nella tabella seguente vengono riportati gli intervalli dei tempi di divisione trovati svolgendo alcune simulazioni di prova, facendo variare diversi parametri della riga di comando.

	ej min	ej	ej max
-n 2 0.2 -g 2 10 -en 0.02 2 0.5 -ej 0.02 2 1	0,0366	0,0398	0,0433
-n 2 0.3 -g 2 10 -en 0.02 2 0.5 -ej 0.02 2 1	0,0269	0,0295	0,0324
-n 2 0.4 -g 2 10 -en 0.02 2 0.5 -ej 0.02 2 1	0,022	0,0243	0,0269
-n 2 0.5 -g 2 10 -en 0.02 2 0.5 -ej 0.02 2 1	0,0205	0,0227	0,0252
-n 2 0.6 -g 2 10 -en 0.02 2 0.5 -ej 0.02 2 1	0,0195	0,0216	0,024
-n 2 0.7 -g 2 10 -en 0.02 2 0.5 -ej 0.02 2 1	0,0194	0,0215	0,0239
-n 2 0.7 -g 2 10 -en 0.06 2 0.5 -ej 0.06 2 1	0,069	0,074	0,0796
-n 2 0.7 -g 2 10 -en 0.05 2 0.5 -ej 0.05 2 1	0,0556	0,0598	0,0646
-n 2 0.7 -g 2 10 -en 0.05 2 0.2 -ej 0.05 2 1	0,0564	0,0606	0,0654
-n 2 0.7 -g 2 10 -en 0.05 2 0.8 -ej 0.05 2 1	0,0541	0,0583	0,0629

Figura 5.10: Prove di inferenza di alcuni `-ej`

Ringraziamenti

A conclusione di questo lavoro, desidero ringraziare tutti coloro che si sono resi disponibili ad aiutarmi a svolgerlo.

I ringraziamenti vanno innanzi tutto agli antropologi, Alessio Boattini e Luca Pagani, per la possibilità di collaborazione che mi hanno offerto. Li ringrazio inoltre per l'interesse dimostrato per gli argomenti trattati, per non essersi mai avviliti (o per non avermelo mai mostrato) quando i risultati ottenuti non erano per niente simili a quelli aspettati e la loro infinita pazienza messa a dura prova in questi mesi di lavoro.
Ringrazio il professor Desalvo, il quale ha accettato con entusiasmo la sfida della collaborazione in una branca che esula dai suoi studi, sacrificando spesso anche del tempo libero nei fine settimana per portare a termine alcune analisi.
Desidero inoltre ringraziare il professor Campanino, per l'importante contributo apportato nella stesura della tesi, per le puntuali correzioni e i preziosi consigli.
Sono riconoscente infine al professor Pettener, per la sua capacità di trasmettere grande passione per l'antropologia, e grazie al quale ho potuto svolgere una tesi di questo tipo.

D'altra parte non posso dimenticare di menzionare coloro a cui sono debitrice, per aver percorso con me un po' del cammino che mi ha portata fino a questo traguardo.
La mia famiglia:
il babbo, per il sostegno incondizionato e silenzioso in ogni scelta fatta;
la mamma, per la tenacia con cui mi ha sempre invitata ad agire e per la sua presenza costante ad ogni piccolo passo compiuto;
Carlotta, sorella e coinquilina ma soprattutto amica sincera che mai si permette di giudicare, fedele ascoltatrice e valida consigliera.
Gli amici e gli scout, che fanno sì che la mia vita non sia dedicata solo a me stessa.

Bibliografia

[1] Camporello L., Baggio R. *Statistica con SPSS. Esercizi e funzioni base.* Egea (2005)

[2] Ross S.M. *Simulation (quarta edizione).* Elsevier Science (2007)

[3] Everitt B.S., Landau S., Leese M. *Cluster Analysis (quarta edizione).* Edward Arnold (2001)

[4] Duran B.S., Odell P.L. *Cluster Analysis: A Survey.* Springer (1974)

[5] Huberty C.J. *Applied Discriminant Analysis.* John Wiley & Sons (1994)

[6] Pollice A. *Statistica multivariata.* Università degli Studi di Bari, Dipartimento di Scienze Economiche e Metodi Matematici

[7] Jobling M., Hollox E., Hurles M., Kivisid T., Tyler-Smith C. *Human Evolutionary Genetics (seconda edizione).* Garland Science (2014)

[8] Campbell N.A. *Principi di Biologia.* Zanichelli (1998)

[9] Pesce G. *Dispense di Zoologia.* Università di L'Aquila, Facoltà di Scienze MM FF NN

[10] *Enciclopedia Treccani*

[11] *Wikipedia*